Population genomics and transcriptomics in the cotton bollworm, *Helicoverpa armigera*

Sue Vern Song

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

April 2018

School of Biosciences

Faculty of Science

The University of Melbourne

Abstract

Helicoverpa armigera is an agricultural pest that causes billions of dollars' worth of damage each year. As *H. armigera* has evolved resistance to insecticides, an understanding of resistance genes will provide useful insights into managing this pest. One approach to identify candidate genes is to scan the genome for signs of strong and recent selective sweeps. This extends the search beyond typical candidate genes (detoxifying enzymes and molecular targets), although a limitation of the approach is that the selective agent causing a sweep may not be an insecticide. Another approach is to use transcriptomics to identify differentially expressed genes between laboratory-selected and unselected cohorts, and between treatments (exposure to an insecticide). Genes that are differentially expressed are good candidates for further investigation.

This thesis begins by estimating some baseline parameters such as nucleotide diversity and the extent of linkage disequilibrium to lay a foundation for detecting selective sweeps in *H. armigera*. In order to detect deviations from the neutral hypothesis, it is necessary to ask 'What does neutral look like in *H. armigera*?' first. Using molecular markers designed to span introns, I survey nine Z-linked loci in females from three Australian populations of *H. armigera*. The choice of Z-linked markers is motivated by the ability to overcome the problem of gametic phase afforded by hemizygosity – sequencing the Z chromosome in the hemizygous sex allows empirical estimates of linkage disequilibrium to be obtained without relying on imputation. I find that *H. armigera* exhibits high levels of nucleotide diversity and rapid decay of linkage disequilibrium. I also uncover a cytochrome P450 gene, *Cyp303a1* that exhibits the hallmarks of strong and recent selection.

Next, I extend the research to non-Australian populations of *H. armigera* and include more Z-linked loci in the experimental design. As Australia is postulated to be the centre of origin for the *Helicoverpa* lineage, it is not clear if the findings of the previous study could be considered 'typical' of *H. armigera* more generally. My assessment of Chinese *H. armigera* populations finds that they too exhibit high levels of diversity and rapid

decay of linkage disequilibrium. A comparison of *H. armigera* from Australia and China reveals that there is generally very little population structure in this species, with the exception of the *Cyp303a1* locus. Using data from a Z chromosome-wide scan performed in an independent set of Australian and Chinese samples, I find that there are 'blocks' of differentiation on the *H. armigera* Z chromosome and propose that the genes found in these regions could be useful markers for discerning between *H. armigera* from different populations. I also identify signs of a population expansion in the history of *H. armigera*.

Finally, I perform comparative transcriptomic analyses to identify differentially-expressed genes between selected and unselected cohorts of *H. armigera* in response to the pyrethroid insecticide fenvalerate. Where genomic scans of selective sweeps fall short in identifying a selective agent, a transcriptomic approach may bridge the gap between genotype and phenotype by validating the biological relevance of candidate loci. To that end, I look for genes that are induced upon exposure to the insecticide, as well as constitutive differences between an unselected strain and a strain from the same genetic background that has been selected for resistance to the insecticide. I also assess the contribution of five detoxification gene families to the resistance phenotype, and find that they are over-represented in the list of differentially-expressed genes. *Cyp303a1* is not found to be amongst the list of cytochrome P450s that are differentially expressed, and I discuss some reasons for this. Finally, I discuss some implications and future directions of the findings presented in this thesis.

Declaration

This is to certify that:

(i) This thesis comprises only my original work towards the Doctor of Philosophy except where indicated in the preface;

(ii) Due acknowledgement has been made in the text to all other material used; and

(iii) This thesis is fewer than 100,000 words in length, exclusive of tables, figure legends, maps, bibliographies and appendices.

Sue Vern Song

Preface

Chapter 2

SVS, John G. Oakeshott (JGO) and Charles Robin (CR) designed the study. Sharon Downes (SD) and Tracey Parker collected and provided samples. JGO and CR provided reagents and equipment. SVS wrote custom scripts for simulations and performed the analyses, advised by JGO and CR. SVS, SD, JGO and CR wrote the manuscript. At least 70% of the work was carried out by SVS. This study has been published in Heredity (2015) 115, 460470; doi:10.1038/hdy.2015.53. Authorisation has been received from co-authors agreeing to the listed publication being included as part of this thesis.

Chapter 3

SVS, JGO and CR designed the study. Craig Anderson (CA) provided data, scripts and advice on the Z chromosome-wide section of the study. Robert Good provided advice on molecular marker design and bioinformatic analyses. Stephen Leslie (SL) provided advice on `Structure` analyses, statistical tests and simulations. Yidong Wu provided samples. JGO and CR provided reagents and equipment. SVS wrote custom scripts for simulations and performed the analyses, advised by SL, JGO and CR. SVS, CA, SL, JGO and CR wrote the manuscript. At least 60% of the work was carried out by SVS. This study has been published in Bulletin of Entomological Research (2018), and is available under First View at http://dx.doi.org/10.1017/S0007485318000081. Authorisation has been received from co-authors agreeing to the listed publication being included as part of this thesis. For inclusion in this thesis, changes have been made to the font size and spacing of the manuscript text, and numbering and placement of captions for aesthetic and indexing reasons.

Chapter 4

This study was carried out in collaboration with members of the Land and Water Division at CSIRO, Black Mountain. Derivation and rearing of *H. armigera* strains (section 4.2.1) were carried out by Peter Hart and Claire Farnsworth. Peter Hart and I set up and performed the assay. I carried out the RNA extractions while Chris Coppin performed the preparation, quantification and normalisation of libraries for RNA-Seq (section 4.2.2 and Appendix A). Read assembly, alignment and preliminary analyses (section 4.2.3) were performed by Stephen Pearce. All other sections in this chapter comprise my own work unless otherwise noted.

Acknowledgements

I would like to start by thanking my supervisors, Charles Robin and John Oakeshott for their support and guidance throughout my PhD – your insightful observations and enthusiasm for your craft have been an invaluable source of ideas, and your mentorship has been instrumental to my growth as a scientist. I am particularly grateful for your patience and understanding over the course of this long journey.

To my committee members, Nancy Endersby and Belinda Appleton – thank you for your advice and guidance. An extra thanks also goes out to Nancy for being a fantastic lab manager and keeping us safe from the wrath of lab auditors.

To members of the Robin lab – thank you for the company, commiseration, laughter and chocolate. Thank you to Alex Fournier-Level, a fount of knowledge on all things R, and Rob Good, who introduced me to the wonderful world of regular expressions, reagent hacking and bioinformatics. Amol, Lisa and Llew – you guys have been great companions inside and outside the lab. Caitlyn and Bec – your discipline and time management skills are truly inspiring and humbling.

Thank you to members of the Oakeshott lab for all their advice and assistance, in particular Claire Farnsworth, Peter Hart, Chris Coppin, Stephen Pearce and David Clarke, and to members of the Hoffmann and Batterham labs, Rahul Rane, Heng Lin Yeap, Ronald Lee, Tinna Yang and Chris Lumb for advice on bioinformatics and molecular techniques.

Last but not least, I thank my family and friends for their unwavering love and support. I am truly fortunate to have all of you in my life. Thank you for the financial and emotional support, time spent channeling messages of encouragement and positive vibes from around the globe, gifts of food and home-cooked meals, and for providing an oft-needed balance of diverse perspectives on life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Helicoverpa armigera is a major lepidopteran pest of agriculture widely occurring throughout Africa, Asia, Europe and Australasia. High polyphagy coupled with resistance to insecticides make it particularly successful at causing significant damage to the yield of economically important crops. In Australia, the losses caused by *H. armigera* along with *H. punctigera* motivated the introduction of transgenic cotton and the development of sustainable management strategies as resistance to DDT and synthetic pyrethroids were detected in the 1970s and 80s (Zalucki *et al.*, 1986; Fitt, 2003). The wide host range and feeding habits of the species are reflected in the list of common names assigned to it including the cotton bollworm, tomato grub, corn earworm and gram pod borer. The recent incursion of *H. armigera* into Brazil (Tay *et al.*, 2013) presents new challenges for resistance management in the Americas. The study of insecticide resistance is relevant for not only utilitarian purposes but also investigating the genetic basis of adaptation. Insecticides are one of the strongest selective agents in the environment, so studying resistance genes in a target pest organism provides unique opportunities to advance our understanding of more general evolutionary processes as the selective agent is known and fitness differences can be observed in a tractable manner (McKenzie and Batterham, 1994; McKenzie, 2000).

1.1 *Helicoverpa armigera* and its closely-related species

The *Heliothinae* subfamily of noctuid moths houses some of the world's most injurious crop pests such as the tobacco budworm, *Chloridea virescens* (formerly *Heliothis virescens*; Pogue (2013)) in addition to the *Helicoverpa* genus. The major heliothine pests owe much of their success to their broad host ranges, high fecundity and high dispersal ability. Apart from the economic interest surrounding pest species, the subfamily acts as a useful model for understanding host range evolution and host adaptation as it contains both specialist and generalist feeders. The phylogenetic position of the *Heliothinae* with respect to other noctuid moths is not fully established due to the large size of the noctuid family, but it is likely that the subfamily forms a monophyletic group (Mitter *et al.*, 1993).

The genus *Helicoverpa* was introduced to house *H. armigera* and its sister species on the basis of morphological structures, but many aspects of heliothine phylogeny were not satisfactorily resolved by morphology due to a lack of informative characters from simplified structures, prompting a study by Cho *et al.* (2008) to reassess the phylogeny using sequences from two nuclear gene regions. Their work in addition to evidence from other studies supports monophyly of the *Helicoverpa* genus, making it a fairly well-defined genus within the heliothine subfamily (Matthews, 1991, 1999; Fang *et al.*, 1997).

Relationships within the genus are less well-defined. Morphological similarities have resulted in misidentification of specimens and crop damage by one species was sometimes misattributed to another, creating difficulties in accurately determining their geographical distributions (Common, 1953; Zalucki *et al.*, 1986). Even with the use of molecular markers, resolution of congeners can be limited by a lack of informative characters. At present, *H. punctigera* is the best candidate for the basal species of the genus (Figure 1.1) and suggests an Australasian origin for the *Helicoverpa* lineage as *H. punctigera* is endemic to Australia (Mitter *et al.*, 1993; Matthews, 1999; Cho *et al.*, 2008). *H. gelotopoeon* is the oldest congener outside of Australia and the most likely source of several species that are endemic to South America. It is unclear where *H. armigera* arose. One hypothesis is that the diversification that gave rise to *H. armigera* and *H. assulta* occurred in Australia because *H. assulta* is closely-related to two Australian endemics, *H. hardwickii* and *H. prepodes* (Matthews, 1999). *H. assulta* is distributed across the Old World but unlike *H. armigera*, it is not viewed as a major pest as it is typically restricted to solanaceous crops.

Interest in the relationship between *H. armigera* and its New World counterpart, *H. zea* has been rekindled in light of *H. armigera*'s incursion into South America. The two were once thought to constitute one species, *Heliothis obsoleta* or *Heliothis armigera* until Hardwick proposed five species groups on the basis of morphological differences: *armigera*, *gelotopoeon*, *hawaiiensis*, *punctigera* and *zea*. Subsequent data from immunological assays and mitochondrial DNA (mtDNA) point to *H. zea* being a closer relative of *H. armigera* than some of the other species within Hardwick's *zea* group such as *H. assulta*

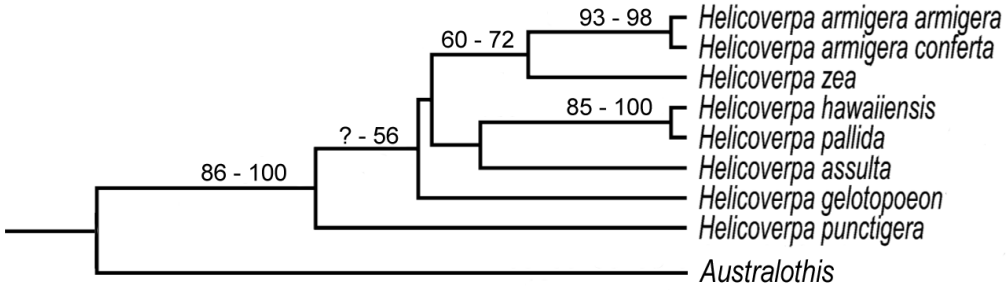


Figure 1.1: Phylogenetic tree of species in the *Helicoverpa* genus, adapted from Cho *et al.* (2008). Numbers represent the range of bootstrap values obtained from two nuclear and one mitochondrial gene regions under maximum parsimony (MP) or maximum likelihood (ML) models. A question mark represents missing data, indicating weak support for the relationship. Note that no bootstrap values were reported where the split between *H. assulta* and the clade containing *H. hawaiiensis* and *H. pallida* occurs in the original figure.

(Greenstone *et al.*, 1991; Behere *et al.*, 2007). At present, *H. zea* and *H. armigera* are thought to be two very closely-related but distinct species that diverged approximately 1.5–2 million years ago (Mya) (Mallet *et al.*, 1993; Mitter *et al.*, 1993; Mitchell and Gopurenko, 2016; Pearce *et al.*, 2017).

There have been reports of hybridisation between *H. armigera* and *H. zea*, and between *H. armigera* and *H. assulta* in the laboratory, which have resulted in fertile offspring (Laster and Hardee, 1995; Wang and Dong, 2001). Although such artificially-induced matings may not be applicable to field populations, the overlapping ranges of *Helicoverpa* species in Australia leave the possibility open. Difficulties with reconciling data from molecular analyses, morphology and matings (the biological species concept) suggest that the species boundary is not always clearly delineated, and that interspecific gene flow is a natural (albeit uncommon) phenomenon. The large amounts of genomic data made accessible by current genome-sequencing technologies are expected to provide more insight into this area in the near future, along with increasingly sophisticated models that include introgression and hybridisation (Marsden *et al.*, 2011; Staubach *et al.*, 2012; Poelstra *et al.*, 2014).

1.2 Recent incursion of *Helicoverpa armigera* into the Americas

The geographic distribution of *H. armigera* was previously restricted to Africa, Australasia, Asia and Europe, with the closely-related *H. zea* having a complementary distribution in the Americas. However, the last five years have witnessed infestations of *H. armigera* in multiple countries across South America (Czepak *et al.*, 2013; Murúa *et al.*, 2014; Arnemann *et al.*, 2016; Sosa-Gómez *et al.*, 2016). Using molecular markers to discriminate between the two species, Tay *et al.* (2013) confirmed that *H. armigera* was indeed responsible for the unusually high rate of attacks and reduced efficacy of control methods in the growing seasons of 2011 to 2013 in Brazil. The origins of the Brazilian populations could not be determined due to the prevalence of haplotypes that were shared with populations found across the Old World. Nevertheless, the implications for Brazilian agriculture are serious due to *H. armigera*'s history of evolving resistance to several classes of insecticides coupled with the lack of an integrated approach for managing resistance in the region. Given its similar ecology to *H. zea* and migratory ability, it seems inevitable that *H. armigera* will expand its range into North America. Where previously the perceived threat was from human-mediated dispersal, deemed isolated events of low frequency and easily eradicated, it now appears more likely that *H. armigera* may enter via natural dispersal pathways resulting in multiple re-invasions (Kriticos *et al.*, 2015). An incursion of *H. armigera* into the United States may require a reexamination of current practices as *H. armigera*'s history of insecticide resistance and global distribution differs from the country's other established heliothine pests.

The potential for fertile offspring arising from hybridisation between *H. armigera* and *H. zea* populations raises concerns that resistance traits from *H. armigera* could easily introgress into *H. zea*. Adaptive introgression has been reported in mice whereby a divergent allele conferring warfarin resistance was detected in *Mus musculus domesticus*, with gene genealogies supporting an introgressive event from *M. spretus* (Song *et al.*, 2011). A second example comes from a study of African malaria mosquitoes, *Anopheles gambiae*

and *Anopheles coluzzii* (previously designated as the M and S forms of the *A. gambiae* complex), two species which differ in their susceptibilities to the insecticides used for malarial control (Weill *et al.*, 2000; Norris *et al.*, 2015). The sympatry of *H. armigera* and *H. zea* in the Americas provides ample opportunity for increased occurrences of hybridisation and introgression, potentially allowing *H. zea* to rapidly acquire resistance genes from *H. armigera*.

1.3 Mechanisms of insecticide resistance

Chemical insecticides continue to be a mainstay of crop protection despite concerns over the sustainability of the practice and their potential toxicity to humans and other species. Biological control carries its own set of environmental concerns and requires considerable effort towards understanding the ecology of both the target organism and control agent. High costs are incurred in the development of a robust yet stable strain of the control agent, so chemical sprays are favoured for their relative ease of use, availability and potential for quick intervention (King and Coleman, 1989). Nevertheless, development of novel insecticides is not trivial and it remains in the best interests of manufacturers to delay the spread of resistance for as long as possible. Given the intense selective pressure and genetic constraints placed on a pest organism, developing resistance appears to be a matter of 'when' rather than 'if'.

Insecticides can be broadly classed by their mode of action based on the physiological functions affected (Insecticide Resistance Action Committee: *http://www.irac-online.org/modes-of-action/*). Examples of mode of action include inhibition of enzymes in the nervous system, blocking of ion channels, inhibition of chitin biosynthesis and synthetic analogues that mimic insect hormones. Mechanisms of insecticide resistance can be broadly described in the following ways: target-site modification, metabolic detoxification, reduced penetration through the cuticle and behavioural modifications. The first two mechanisms are the most common whereby resistance occurs due to decreased sensitivity of the target and/or increased detoxification of the insecticide.

1.3.1 Target-site insensitivity

Conventional targets of insecticides include enzymes or receptors involved in the insect nervous system and ion channels. Resistance arises from modifications to the target sites, resulting in a failure of the insecticide to bind. Target-site insensitivity usually causes very high levels of resistance compared to that seen in metabolic detoxification. Additionally, one site may be targeted by multiple classes of insecticides, so target-site modifications can cause cross-resistance, severely limiting the options available for pest control. These modifications typically take the form of point mutations that cause amino acid substitutions, and the same mutations can be observed in multiple species as the target sites are highly conserved.

At the *resistance to dieldrin* (*Rdl*) locus for instance, the same replacements (alanine to serine or glycine at position 302) were consistently found in a wide range of insect species including *D. melanogaster*, *D. simulans*, beetles, whitefly, green peach aphid, mosquito and cockroach (Anthony *et al.*, 1998; ffrench Constant *et al.*, 1998). In most species, *Rdl* occurs as a single-copy gene that codes for a γ-aminobutyric acid (GABA) receptor subunit targeted by cyclodiene insecticides. GABA is a major inhibitory neurotransmitter present in both vertebrates and invertebrates. Blockage of the GABA-gated chloride ion channel causes lethality through hyperexcitation of the nervous system (Bloomquist, 1993). However, the situation in lepidopterans may be further complicated by the presence of three paralogues of *Rdl* which could either delay or hasten the development of resistance (Heckel, 2009).

Another target is the enzyme acetylcholinesterase, which is responsible for degrading the neurotransmitter acetylcholine. Organophosphate and carbamate insecticides inhibit the enzyme's ability to hydrolyse acetylcholine, leading to prolonged stimulation. Two patterns of resistance have been identified: in Pattern I, insensitivity to carbamates is greater than to organophosphates whereas in Pattern II, there are similar levels of insensitivity to both classes of insecticides (Russell *et al.*, 2004). Both patterns have been

observed in lepidopterans, including *H. armigera* (Heckel, 2009). Higher Diptera have one *acetylcholinesterase* (*Ace*) gene while other insects including *H. armigera* have two (Ren *et al.*, 2002).

Pyrethroids and DDT target the voltage-gated sodium channel, causing hyperexcitability, paralysis and eventual death. *knockdown resistance* (*kdr*) is the most common form of resistance against these insecticides, while a second trait, designated super-*kdr* has been found to confer even greater levels of resistance. Studies of *kdr* and super-*kdr* houseflies found that resistance was the result of modifications to the *Vssc1* locus, which was the orthologue of the Drosophila *para* sodium channel in houseflies. Further evidence for the role of *para* in pyrethroid resistance came from mapping studies which showed close genetic linkage between *kdr* and the *para* genes in tobacco budworm and German cockroach (Davies *et al.*, 2007). The *para* locus has also been used to explore genetic variation in *H. armigera*, although no association was found between haplotype frequency and fenvalerate resistance (Stokes *et al.*, 1997). However, the authors noted that this result is consistent with a reported increase in resistance mechanisms due to metabolic detoxification rather than target-site insensitivity (Gunning *et al.*, 1991).

1.3.2 Metabolic detoxification

The relatively small number of insecticide targets have facilitated our understanding of target-site insensitivity, providing valuable insights into the mechanisms underlying resistance as well as examples of parallel or convergent evolution in action. Characterisation of metabolic detoxification, on the other hand, is encumbered by the large number and variety of enzymes involved in an organism's physiological processes. Compared to target-site modifications which are well-conserved across different insect orders, the enzymes involved in metabolic detoxification can vary widely from species to species. Resistance can arise from gene amplifications, transcriptional changes or mutations in coding sequences of the genes encoding these enzymes, thus providing the insect with a wide range of genetic options for evolving resistance, including cross-resistance to multiple classes of insecticides (Li *et al.*, 2007; Oakeshott *et al.*, 2013). Characterisation of resistance due to metabolic

detoxification has typically revolved around members of gene families involved in phase I, II and III metabolism, detoxification, sequestration and/or removal of xenobiotics (Xu *et al.*, 2005; Dawkar *et al.*, 2013).

Phase I metabolising enzymes mainly involve the cytochrome P450s (CYPs), a superfamily of enzymes with broad substrate specificity. Overexpression and amplification of CYPs have been implicated in resistance to several classes of insecticides (Carino *et al.*, 1994; Zhao *et al.*, 1996; Daborn *et al.*, 2001; Edi *et al.*, 2014; Puinean *et al.*, 2010; Lin *et al.*, 2013; Xu *et al.*, 2016) including chlorantraniliprole, a new class of insecticide that targets the insect ryanodine receptors (Bassi *et al.*, 2009). CYPs also have other roles in insects such as ecdysteroid metabolism, pheromone metabolism and cuticle formation (Feyereisen, 2005) which could potentially contribute to non-detoxificative mechanisms of resistance. Phase I serves to reduce the toxicity of the insecticide while Phase II converts hydrophobic toxic compounds into hydrophilic products to facilitate excretion. Phase II metabolising or conjugating enzymes include superfamilies such as sulfotransferases, UDP-glycosyltransferases (UGTs), glutathione S-transferases (GSTs) and carboxyl/cholinesterases (CCEs). Gene amplification and altered DNA methylation in CCEs have been implicated in resistance to organophosphates, carbamates and pyrethroids (Devonshire and Moores, 1982; Field *et al.*, 1988, 1989) while GST gene amplification and overexpression have been implicated in resistance to DDT, organophosphates and pyrethroids (Grant and Hammock, 1992; Prapanthadara *et al.*, 1993; Huang *et al.*, 1998; Vontas *et al.*, 2001; Ortelli *et al.*, 2003). Less is known about the contribution of UGTs to insecticide resistance, but they could have a role in detoxifying DDT, carbamates, pyrethroids and chlorantraniliprole (Pedra *et al.*, 2004; Vontas *et al.*, 2005; Silva *et al.*, 2012; Li *et al.*, 2016).

Finally, Phase III transporters act to remove the conjugated xenobiotics from the cell. The superfamily of ATP-binding cassette transporters (ABCs) which includes P-glycoproteins are capable of either importing or exporting a broad range of substrates, though they typically function as exporters in eukaryotic cells. Apart from their roles as transporters, they also function as receptors, ion channels and regulators of channels (Kuchler, 2011;

Dermauw and Van Leeuwen, 2014). In insects, ABCs have been implicated in resistance to several classes of chemicals including carbamates, organophosphates and pyrethroids (Lanning *et al.*, 1996; Srinivas *et al.*, 2004; Epis *et al.*, 2014). ABCs have also been implicated in lepidopteran resistance to *Bacillus thuringiensis* (*Bt*) toxins. In the model proposed by Gahan *et al.* (2010), ABCC2 is hypothesised to be a binding target for the Cry1Ab and Cry1Ac toxins, facilitating formation of the toxin pore in the membrane. Resistance arises from an inactivating mutation that leads to loss of the binding site. The C and G subfamilies of ABCs have particularly high frequencies of duplications (Labbé *et al.*, 2011), which suggests that they could have key roles in detoxification.

1.3.3 Reduced penetration through the cuticle

Target-site insensitivity and metabolic detoxification can be supplemented by reduced penetration of the insecticide through the cuticle. Mechanisms such as thickening of the cuticle or changes in cuticular hydrocarbon content can reduce the amount of toxin that comes into contact with the insect's internal organs and/or allow more time for the insecticide to be detoxified and excreted (Pedrini *et al.*, 2009; Wood *et al.*, 2010). A study involving DDT-resistant strains of *D. melanogaster* reported that *Cpr72Ec* (CG4784) and *Lcp1* (CG11650) were upregulated relative to the susceptible strain (Qiu *et al.*, 2013). Both genes are thought to be involved in chitin-based cuticle formation but their roles are not fully understood.

The main contribution of reduced penetration is to delay the absorption and distribution of toxins in the insect haemolymph. Apart from the active ingredient, the carrier molecule or solvent used can also affect the absorption rate as these chemicals also interact with the constituents of the insect cuticle, so this has implications for pesticide formulations (Sawicki and Lord, 1970). Overall, reduced penetration through the cuticle in itself appears to have only marginal effects on resistance. However, its role as an initial barrier or delaying factor may significantly increase resistance in populations where target-site insensitivity and/or detoxification mechanisms are already present (Sawicki, 1970; Hama *et al.*, 1987; Noppun *et al.*, 1989).

1.3.4 Behavioural modifications

Resistance due to behavioural modifications is poorly-characterised presumably due to the complex genetic architecture underlying such traits, plus the challenge of defining the scope of the behavioural response to be assayed in the first instance. A contentious issue is whether or not observed differences in behaviour are a result of heritable changes (in chemosensory receptors, for instance), aversion due to learning, or simply an artefact of weaknesses in the study design (Zalucki and Furlong, 2017). Nevertheless, the study of behavioural modifications remains relevant as it affects decisions about where and how pesticides should be applied in order to maximise contact with the target pest. Further study into the physiological mechanisms underlying behavioural resistance could help inform strategies for integrated pest management (Wang *et al.*, 2004; Nansen *et al.*, 2016).

1.4 Insecticide resistance in *Helicoverpa armigera*

The introduction of an integrated resistance management (IRM) strategy in 1983 arose as a response to the development of pesticide resistance in the Australian cotton industry. Resistance to DDT in *H. armigera* had already been documented in the Ord River Valley of Western Australia in the late 1960s. By the time of the Ord industry's collapse in 1973, DDT resistance had also been reported in the eastern states. Detection of endosulphan and pyrethroid resistance followed, leading to alternations in the use of different chemical classes for control (Wilson, 1974; Forrester *et al.*, 1993; Fitt, 1994). Studies of the mechanisms underlying pyrethroid resistance in Australian *H. armigera* revealed that target-site insensitivity was primarily responsible for the high order of resistance in populations sampled before implementation of the IRM (Gunning *et al.*, 1991, 1995). By contrast, metabolic detoxification and reduced penetration through the cuticle were predominant in populations sampled after the restrictions on pyrethroid use, with resistance occurring at a lower order.

Studies of non-Australian *H. armigera* populations indicate that similar mechanisms are present. Pyrethroid control failures have been reported since *H. armigera* was detected in Brazil where it is now a major pest of soybean and cotton (Durigan *et al.*, 2017). Nerve insensitivity reminiscent of *kdr* has been observed in field populations from Asia (Ahmad *et al.*, 1989; McCaffery *et al.*, 1997). Pyrethroid resistance due to reduced penetration has been documented whereby recovery of ^{14}C-labelled insecticide from the cuticle of resistant strains was higher than from susceptible strains (Gunning *et al.*, 1991; Ahmad *et al.*, 2006). Several reports have linked altered amino acid sequences and/or upregulation of CYPs and CCEs to pyrethroid and organophosphate resistance although further study is needed to determine the metabolic capabilities of these enzymes (Srinivas *et al.*, 2004; Brun-Barale *et al.*, 2010; Wu *et al.*, 2011; Yang *et al.*, 2013; Xu *et al.*, 2016). A novel resistance mechanism has been described in *H. armigera* whereby a chimeric gene that arose from unequal crossing-over between two functional parental genes resulted in an enzyme capable of metabolising fenvalerate efficiently (Joußen *et al.*, 2012). The presence of this chimera, *Cyp337b3* and its allelic variants has been documented in populations of *H. armigera* from Australia, China, Pakistan and Brazil (Rasool *et al.*, 2014; Han *et al.*, 2015; Joußen and Heckel, 2016; Durigan *et al.*, 2017).

Expression of P-glycoproteins (Pgp) has been implicated in pyrethroid, organophosphate and abamectin resistance while ABCA2 and ABCC2 have been linked to *Bt* resistance in *H. armigera* (Srinivas *et al.*, 2004; Xiao *et al.*, 2014; Tay *et al.*, 2015; Xiang *et al.*, 2017). Pgp's role in resistance appears to be mainly that of a transporter involved in xenobiotic efflux. This contrasts with ABCA2 and ABCC2 which act as binding targets for *Bt* toxins, and mutations in the protein sequences confer the high levels of resistance typically observed in target-site insensitivity. Interestingly, mutations in ABCC2 that conferred resistance to *Bt* toxins appeared to increase susceptibility to abamectin (Xiao *et al.*, 2016). Such instances of negative cross-resistance may increase the sustainability of pesticide application and provide a temporary respite in the race between humans and their insect pests.

Table 1.1 outlines the findings and limitations of studies that have investigated the ge-

netic basis of resistance in *H. armigera*, in particular, resistance to pyrethroids. In most cases, overexpression in resistant strains of *H. armigera* is cited as evidence of a gene's contribution to the resistance phenotype. However, the different genetic backgrounds of the resistant and susceptible strains are often not taken into consideration. A more informative approach would be to compare expression profiles after controlling for the genetic background of both resistant and susceptible strains and/or assess gene function through enzyme activity assays or targeted gene replacement.

Gene	Insecticide class	Findings	Limitations
Esterases (Clade 1, 14, 16 17)	Pyrethroids, organophosphates	Baculovirus expression of genes produced isozymes with activity against insecticides (Teese *et al.*, 2013)	Alleles have not been characterised
CYP337B1, CYP337B3	Pyrethroids	Linkage analysis maps to this locus (Wee *et al.*, 2008), CYP337B3 metabolises fenvalerate to a non-toxic product (Joußen and Heckel, 2016)	Little correlation between resistance and some alleles of CYP337B3
CYP4G8	Pyrethroids	Overexpression in resistant strain (Pittendrigh *et al.*, 1997)	Functional evidence and enzyme activity have not been established
CYP4S1	Pyrethroids	Overexpression in resistant strain (Wee *et al.*, 2008)	Functional evidence and enzyme activity have not been established
CYP6B2, CYP6B6, CYP6B7	Pyrethroids	Induction by permethrin (Wang and Hobbs, 1995), overexpression in resistant strain (Ranasinghe and Hobbs, 1998; Ranasinghe *et al.*, 1998)	Linkage analysis of a pyrethroid-resistant strain did not map to these genes (Grubor and Heckel, 2007), although they could be responsible for resistance in other strains
CYP9A12	Pyrethroids	Activity against fenvalerate using yeast expression system (Yang *et al.*, 2008), induction by deltamethrin (Zhou *et al.*, 2010a), overexpression in resistant strain (Brun-Barale *et al.*, 2010; Xu *et al.*, 2016)	Functional evidence (e.g. targeted gene replacement) has not been established in *H. armigera*
CYP9A14	Pyrethroids	Activity against fenvalerate using yeast expression system (Yang *et al.*, 2008), overexpression in resistant strain (Brun-Barale *et al.*, 2010)	Functional evidence (e.g. targeted gene replacement) has not been established in *H. armigera*
CYP4L5, CYP4L11, CYP4M6, CYP4M7, CYP6AE11, CYP332A1	Pyrethroids	Overexpression in resistant strain (Brun-Barale *et al.*, 2010; Xu *et al.*, 2016)	Functional evidence and enzyme activity have not been established
Z-linked factor	Pyrethroids	Linkage analysis of endosulphan-resistant strains found sex-linked inheritance (Daly and Fisk, 1998)	Genetic basis not fully characterised

Table 1.1: Findings and limitations of studies that have investigated the genetic basis of pyrethroid resistance in *H. armigera*

1.5 Genetic variation in *H. armigera*

Strategies for IRM are based on the premise that the increase in frequency of resistance alleles in the field can be slowed by altering the selective pressure resulting from the use of insecticides (Roush and Tabashnik, 1991). The study of genetic diversity allows us to predict the spread of resistance as patterns of variation can be used to infer population structure and dispersal. The success of *H. armigera* as an agricultural pest and its resistance status motivated the characterisation of gene flow in this species in Australia (Gunning and Easton, 1989; Daly, 1993). Genetic variation in *H. armigera* has been explored via allozymes (Daly and Gregg, 1985; Fisk and Daly, 1989; Nibouche *et al.*, 1998), RAPDs (Zhou *et al.*, 2000), mitochondrial DNA (Behere *et al.*, 2007), microsatellites (Scott *et al.*, 2003, 2004; Ji *et al.*, 2005) and exon-primed intron-crossing (EPIC) markers (Tay *et al.*, 2008). Stokes *et al.* (1997) explored allelic diversity in Australian *H. armigera* using the *para* sodium channel gene and noted the high levels of variation present in such a small region (at least 16 alleles in a 250-bp region), although no associations between allele frequency and insecticide resistance were observed. Overall, there appears to be little evidence for population structure in *H. armigera*. In cases where it has been reported, there are typically no strong geographical patterns, and F_{ST} values are low even among populations separated by geographical barriers (Daly and Gregg, 1985; Endersby *et al.*, 2007). The ability of *H. armigera* to undertake long-distance migration (Fitt, 1989) and the mating compatibility of insects raised on different host plants (Vijaykumar *et al.*, 2007) suggest a large effective population size and high gene flow even if populations were structured due to feeding preferences. The presence of shared haplotypes in collections from different continents supports the idea of *H. armigera* as a panmictic species (Behere *et al.*, 2007).

Various limitations are associated with the different markers. The dominant nature of RAPDs makes heterozygotes indistinguishable from homozygotes, and allozymes do not reflect variation at the nucleotide level. Microsatellites in Lepidoptera suffer from high frequencies of null alleles and/or associations with transposable elements (Zhang, 2004; Endersby *et al.*, 2007; Weeks *et al.*, 2010). The occurrence of null alleles is interesting

in itself as it is suggestive of sequence polymorphisms resulting in the inability of primers to anneal or failure in yielding observable products due to large insertions. While EPIC markers are not immune to these failures, they have been successfully employed to demonstrate polymorphisms in single-copy genes without the occurrence of allele dropouts (Tay *et al.*, 2008) and allow for extending the analyses to closely-related species. EPIC markers thus form the backbone of the sequence data gathered in this thesis.

1.5.1 F_{ST}

F_{ST} is a widely-used statistic for estimating levels of genetic differentiation between populations based on the variance in allele frequency between populations or subpopulations. Low values of F_{ST} imply that allele frequencies are similar between populations, and that little population structure exists. If a particular locus is under selection, it may result in an F_{ST} value that is higher compared to other neutrally-evolving loci. F_{ST} can be used in this manner to identify regions under selection in genome-wide scans (reviewed in Holsinger and Weir, 2009).

1.6 Neutral theory and tests of selection

The study of genetic variation provides empirical measures to assess the effects of mutation, selection and drift. These three forces interact to shape the genomic landscape of populations over time, and their effects are intertwined. Mutation increases genetic variation in a population by introducing changes or novel elements into the genome, whereas drift reduces genetic variation through a random loss of alleles as a consequence of randomly sampling gametes from one generation to the next. Selection can act to either increase or decrease the amount of genetic variation relative to the previous generation; for instance, purifying selection decreases genetic variation by nudging optimal genotypes towards fixation, whereas balancing selection maintains advantageous polymorphisms in a population. The neutral theory provides a useful framework to disentangle the effects of these forces.

Kimura's neutral theory of molecular evolution put forward the idea that most mutations at the molecular level are selectively neutral. Tests of selection therefore revolve around predictions from the neutral theory as the null hypothesis. Some of these predictions include expectations about how mutations are distributed within populations (the frequency spectrum), linkage disequilibrium and haplotype structure (Nielsen, 2005). To demonstrate that a sequence is under selection, the null hypothesis must be rejected i.e. the changes observed in allele frequency must be incompatible with the effects of mutation and drift alone (Kimura, 1983).

1.6.1 Effective population size, N_e

The impact of drift is not limited to neutral variation; beneficial alleles may be lost while weakly deleterious mutations may reach fixation purely through chance (Duret, 2008). As genetic drift is a consequence of sampling error, population size is an important parameter for many of the underlying assumptions in the standard neutral model. In an ideal population, every parent has an equal chance of being the parents of any individual in the next generation (Crow and Kimura, 1970). In reality, not all individuals in a population will contribute equally to the next generation. Effective population size, N_e thus differs from the population census size and uses the concept of an ideal population to estimate the rate of change in a finite population due to genetic drift (Charlesworth, 2009). Smaller populations are more heavily influenced by drift, which results in greater fluctuations in allele frequency even when sequences are evolving neutrally. The standard neutral model assumes random mating, no population subdivision, constant population size and discrete generations. Factors such as population structure, bottlenecks and inbreeding have the effect of reducing N_e, so the population models used in formulating the null hypothesis should consider these factors.

1.6.2 Coalescence

The coalescent is the lineage of alleles in a sample traced backward in time to their common ancestor. Under a neutral model, sampled lineages are 'randomly picking their parents as we go back in time' (Rosenberg and Nordborg, 2002). Coalescence occurs whenever two lineages pick the same parents. Larger populations will have a slower rate of coalescence as there are more parents to choose from. The coalescent is a useful model for formulating expectations about how variation will be distributed in a sample of allelic sequences. One of these expectations is that on average, nucleotide diversity (π) for a sample will be roughly equal to the population scaled mutation rate, or

$$\pi = 4N\mu$$

where μ is the neutral mutation rate per generation and N is the population size. By simulating the genealogy of the sample backwards in time then adding mutations along the branches of the tree, we can come up with some expectations about how often a mutation will be observed $1/n$ times, $2/n$ times, and so on in a sample of n sequences. The coalescent process can be simulated to produce a random genealogy each time, so multiple iterations of this process can be used to derive the allele or site frequency spectrum.

1.6.3 The frequency spectrum

The frequency spectrum describes the distribution of the frequencies of segregating sites in a sample. For instance, the following sample of five sequences containing six segregating sites

Sequence 1: – – – – – – x – – – – – x – – – –

Sequence 2: – – x – – – – – – – – – x – – – –

Sequence 3: – – – – – – – – x – – – x – – – –

Sequence 4: – – x – – – – – – – – – x – – – –

Sequence 5: – – – – x – – – x – – – – – – – x –

would result in a frequency spectrum of (3/6, 2/6, 0, 1/6) as there are three instances where a segregating site is seen once, two instances where it is seen twice, and only one instance where it is seen four times. The frequency spectrum is sensitive to changes in population size, so deviations from the expected distribution could be interpreted as either due to demography or selection.

1.6.4 Linkage disequilibrium

Apart from a skewed frequency spectrum, linkage disequilibrium (LD) and haplotype structure can also provide evidence of selection. LD refers to a non-random association of alleles at two or more loci. It describes a situation where the observed frequency of a particular haplotype differs significantly from an expected outcome, given the random formation of haplotypes and the allele frequencies in the population. For instance, if two alleles are present at locus 1 and locus 2, each at 50% frequency, the expectation is that four haplotypes would be observed, each at approximately 25% frequency. If the ratios differ significantly from the expected values or fewer than four haplotypes are observed, LD can be said to occur. If there were no selection acting, the expectation is that equilibrium would eventually be restored through recombination and random mating in the population (Hill and Robertson, 1968). Levels of LD vary between species and even between genomic regions within a species due to factors such as recombination rates, population sizes and demography. Some expectation of what constitutes 'typical' levels of LD under neutral evolution therefore needs to be formulated prior to testing for selection. The metric used in this study relates to the distance (number of base pairs) at which the correlation between alleles decays – the plots of LD decay curves provide a convenient visual representation to detect atypical patterns when comparing different genomic regions.

1.6.5 Selective sweeps

A selective sweep describes a process whereby a beneficial mutation arises in a population and 'sweeps' away the variation at linked neutral sites as it increases in frequency (Figure

1.2). An increase in the fitness of individuals carrying the beneficial mutation leads to an increased prevalence of linked neutral sites that hitchhike along with the beneficial mutation, causing variation to be reduced in adjacent regions. Variation within a species is also reduced relative to variation between species, and this feature can be used to distinguish between a selective sweep and negative (purifying) selection as the latter reduces both intra- and interspecific variation.

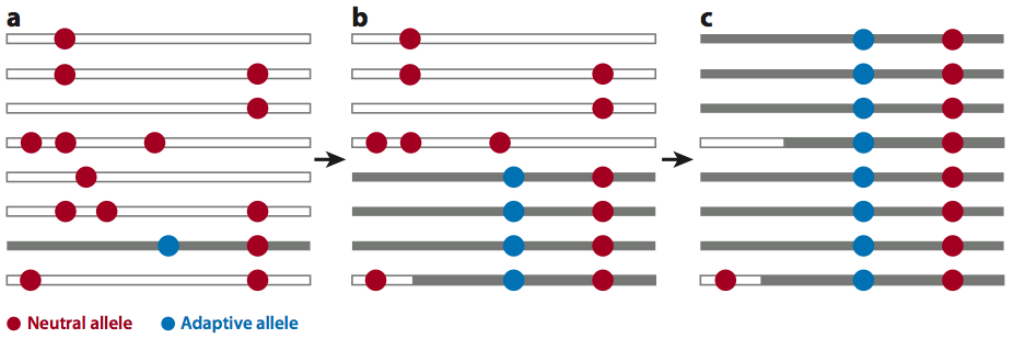


Figure 1.2: Different stages of a selective sweep, adapted from Kelley and Swanson (2008). A beneficial allele arises in a population (a). As it slowly increases in frequency (b), neutral mutations linked to the beneficial mutation also increase in frequency. In an incomplete selective sweep, several ancestral alleles are present along with the swept allele. As the swept allele approaches fixation (c), it leaves behind a signature of reduced variation in the region surrounding the beneficial mutation.

A selective sweep distorts the coalescent process and frequency spectrum because individuals do not have an equal chance of being the parents of any individual in the next generation. When a beneficial mutation causes fitness differences between individuals, some individuals will make a bigger contribution to the next generation than others, leading to an increase in the proportion of alleles that are identical by descent. A selective sweep also increases the prevalence of LD and long haplotypes due to the increased co-occurrence of the beneficial mutation and linked neutral sites. Tests of selection draw on detecting patterns of variation and LD that deviate from the neutral model (reviewed in Nielsen, 2005).

1.6.6 Tajima's *D*

This study primarily uses Tajima's D as the test statistic to detect skewness in the frequency spectrum. Tajima's D is based on the premise that nucleotide diversity calculated from the average number of pairwise differences between sequences, π, should be roughly equal to that calculated from the expected number of segregating sites (Watterson's estimator, θ_W) in the sample (Tajima, 1989). Tajima's D is described by the equation

$$D = \frac{\pi - \theta_W}{\sqrt{\hat{V}(\pi - \theta_W)}}$$

where $\hat{V}$ is the variance of the estimate. New mutations that arise after a sweep have the effect of producing an excess of rare (singleton) alleles in a population. Rare polymorphisms (many segregating sites at low frequency) increase the value of θ_W relative to π. However, the signature of a sweep also resembles that of a population that is growing after a recent bottleneck, so a negative Tajima's D could be due to either a selective sweep or population expansion. A positive Tajima's could be a sign of balancing selection or population contraction. The competing hypotheses (selection or demography) can be distinguished by examining multiple loci – if the signature is limited to a particular locus, it suggests that selection is acting on that locus whereas if the signature is seen across multiple loci, a demographic explanation is more parsimonious.

1.7 From population genetics to population genomics and transcriptomics

A population genetic approach provides a framework for us to understand how patterns of variation are affected by mutation, selection and drift. Genome sequencing projects of model organisms have expanded our understanding of molecular evolution, providing us with bigger data sets for hypotheses-testing and statistical analyses. Different strategies can be employed to identify new candidate genes for traits of interest. Genome-wide

association studies (GWAS) analyse the significance of differences in allele frequencies to associate a particular variant with the phenotype of interest. Panels of inbred strains such as the *Drosophila melanogaster* Genetic Reference Panel (DGRP) are a valuable resource because differences between the strains capture the naturally-occurring segregating variants, while the homozygosity of each strain allows reproducibility and increased accuracy in estimating genotypic parameters (Mackay *et al.*, 2012). Another strategy involves identifying regions which appear to be targets of positive selection by performing genome-wide scans to look for regions that deviate from expectations under the neutral model. While genomic scans for selection offer an exciting and unbiased way of finding novel candidates, it is difficult to ascribe genomic signatures to a selective agent without some functional validation of a genotype-to-phenotype association. Transcriptome-wide scans offer a complementary way of identifying resistance genes and validating the functional significance of candidate loci. By looking for differential gene expression between populations subjected to varying degrees of exposure to an insecticide, we can identify candidates that are biologically relevant in responding to perturbations caused by the insecticide.

One approach is a comparative analysis of susceptible and resistant individuals or populations. The problem of comparing strains with inherently different genetic backgrounds can be mitigated by starting with a single strain, subjecting a subset to a selection regime, then maintaining the selected (resistant) and unselected (susceptible) cohorts in parallel for a number of generations. Genes that show differential expression between the selected and unselected cohorts and/or upon exposure to the insecticide are strong candidates for resistance. One of the most common means of acquiring resistance is through metabolic detoxification of the insecticide. Changes in expression levels of particular genes in the presence of an insecticide can point to candidates for resistance. Further, a substrate is able to affect the activity of the enzymes that can metabolise it, so genes that are induced upon exposure to an insecticide potentially play a role in resistance. Additionally, differential expression between the selected and unselected cohorts in the absence of the insecticide can be used to identify genes that have different basal levels of constitutive expression.

1.8 Thesis aims and outline

As a target pest of insecticides, the genome of *H. armigera* is expected to show signs of selection and patterns of variation that deviate from neutrality. New genomic technologies allow population genetic studies to identify loci potentially exhibiting signs of positive selection or selective sweeps. This thesis aims to lay a foundation for future genome-wide scans of selective sweeps in *H. armigera* by characterising some baseline population genomic parameters, beginning with the detection and characterisation of variation using appropriate marker loci. In Chapter 2 (Paper 1), I develop EPIC markers and use them to estimate levels of diversity and characterise patterns of LD in Australian *H. armigera*. Chapter 3 (Paper 2) extends the use of these markers to characterise variation and LD in non-Australian populations of *H. armigera*, and to assess the extent of population structure between these inter-continental samples. In Chapter 4, I use a transcriptomic approach to identify candidate resistance genes by looking for differential gene expression between a strain that has been selected for resistance to fenvalerate, a synthetic pyrethroid, and an unselected strain of the same genetic background. I then discuss some of the implications of these findings and their utility.

Chapter 2

High nucleotide diversity and limited linkage disequilibrium in *Helicoverpa armigera* facilitates the detection of a selective sweep

2.1 Introduction

The primary goal of this paper is to estimate some baseline parameters to inform future study designs that intend to use a population genomic approach. In genome-wide association studies and scans for selective sweeps, for instance, estimates of nucleotide diversity and linkage disequilibrium provide useful guidelines or constraints that dictate minimum marker density as well as the number of alleles, individuals and/or populations to sample. Having a baseline also allows us to detect deviations from the neutral hypothesis and to identify regions under selection. Prior to this study, population genetic studies of *H. armigera* were typically focused on measures of heterozygosity for the purpose of characterising population structure. Few studies employed measures of diversity with the express aim of laying a foundation for genome-wide scans of selective sweeps in this species. By quantifying nucleotide diversity, π, and the extent of linkage disequilibrium in an important pest organism, this study addresses a gap in the literature for *H. armigera* in the context of insect genomics. Future studies can also draw on these estimates to identify candidate regions under selection for a trait of interest, such as insecticide resistance. This paper thus provides a novel contribution to the field of population genomics and applied pest research.

2.2 Paper 1

Heredity (2015) 115, 460–470

www.nature.com/hdy

ORIGINAL ARTICLE

High nucleotide diversity and limited linkage disequilibrium in *Helicoverpa armigera* facilitates the detection of a selective sweep

SV Song[1], S Downes[2], T Parker[2], JG Oakeshott[3] and C Robin[1]

Insecticides impose extreme selective pressures on populations of target pests and so insecticide resistance loci of these species may provide the footprints of 'selective sweeps'. To lay the foundation for future genome-wide scans for selective sweeps and inform genome-wide association study designs, we set out to characterize some of the baseline population genomic parameters of one of the most damaging insect pests in agriculture worldwide, *Helicoverpa armigera*. To this end, we surveyed nine Z-linked loci in three Australian *H. armigera* populations. We find that estimates of π are in the higher range among other insects and linkage disequilibrium decays over short distances. One of the surveyed loci, a cytochrome P450, shows an unusual haplotype configuration with a divergent allele at high frequency that led us to investigate the possibility of an adaptive introgression around this locus.
Heredity (2015) **115,** 460–470; doi:10.1038/hdy.2015.53; published online 15 July 2015

INTRODUCTION

New genomic technologies allow population genetic studies to move beyond questions of migration and population structure generally to those that identify loci within the genome that exhibit extreme gene flow or population structure or other signs that may be interpreted as selection. One strategy to identify potential insecticide-resistance loci is to seek genomic regions that appear to exhibit the characteristics of positive selection such as extended linkage disequilibrium (LD), reduced nucleotide variation and increased proportions of rare variants in the frequency spectra (Nielsen, 2005). These parameters are expected to vary between different populations and different regions in the genome due to the interplay between drift, recombination, mutation and selection. Consequently, some inquiry into what constitutes the baseline population genomics parameters of a species is required before deviations from neutral expectations can be detected.

A genome-wide survey of molecular variation within the model lepidopteran, *Bombyx mori*, reported that LD decayed over very short distances, with the implication that selective sweeps would be limited to small regions (Xia *et al.*, 2009). Signals of selection were detected at 1041 regions of which 354 were protein-coding genes. These were deemed good candidates for domestication genes, including those involved in silk production, as there has been recent strong selection for such traits. It is reasonable to propose that, in pesticide-resistant organisms where extremely strong selection is exerted on natural populations, similar approaches may identify new candidate resistance genes.

Helicoverpa armigera is a significant lepidopteran pest of agriculture throughout Africa, Asia, Europe and Australia. High polyphagy coupled with an ability to rapidly evolve resistance to insecticides make it responsible for damage to crops estimated at >US$2 billion annually. Resistance to insecticide sprays in *H. armigera* drove the introduction of insecticidal transgenic cotton to Australia and Asia. The recent incursion of *H. armigera* into Brazil (Tay *et al.*, 2013) also threatens agricultural productivity in the New World. Population genomics approaches can characterize past and present population structure throughout the species range and identify adaptive loci such as those that confer resistance to insecticides.

Helicoverpa is a well-defined genus within the heliothine subfamily of noctuid moths where its monophyly is strongly supported by morphology and molecular characterization (Matthews, 1999; Cho *et al.*, 2008). Within the genus, however, relationships between species are less clear often due to morphological similarities. For instance, crop damage by *H. armigera* in Australia is sometimes misattributed to *H. punctigera* and *vice versa* (Zalucki *et al.*, 1986). *H. armigera* and its New World counterpart, *H. zea* were once thought to constitute one cosmopolitan species but Hardwick (1965) placed them into separate species groups when he distinguished five species groups on the basis of penis structure: *armigera*, *gelotopoeon*, *hawaiiensis*, *punctigera*, and *zea*. Subsequent data from immunological assays and mitochondrial DNA sequence analyses suggested that *H. zea* is more closely related to *H. armigera* than some other species within the *zea* group such as *H. assulta* (Mitter *et al.*, 1993; Behere *et al.*, 2007). Hybridization between *H. armigera* and *H. assulta* and between *H. armigera* and *H. zea* is possible in the laboratory (Laster and Hardee, 1995; Wang and Dong, 2001). *H. armigera* and *H. assulta* are sympatric, whereas the geographical distributions of *H. zea* and *H. armigera* were not

[1]Department of Genetics, University of Melbourne and Bio21 Institute, Melbourne, Victoria, Australia; [2]Agriculture Flagship, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Narrabri, New South Wales, Australia and [3]Land and Water Flagship, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australian Capital Territory, Australia
Correspondence: Dr C Robin, Department of Genetics, University of Melbourne, Bio21 Institute, Room 267, Parkville, Victoria 3010, Australia.
E-mail: crobin@unimelb.edu.au
Received 9 February 2015; revised 1 May 2015; accepted 6 May 2015; published online 15 July 2015

thought to overlap in the field until the recent incursion of *H. armigera* into Brazil.

Previous molecular population genetic studies of *H. armigera* used various markers and usually aimed to characterize population structure rather than identify loci under selection. An allozyme study of 12 *H. armigera* populations dispersed throughout Australia suggested limited population structure ($FST = 0.01$; Daly and Gregg, 1985). Similarly, analysis of mitochondrial sequences reveals minimal differentiation among global samples with most of the variation distributed throughout the species range (Behere *et al.*, 2007). A survey of eight microsatellite loci revealed that *FST* among Australian populations was 0.003 after the high frequency of null alleles (>10%) was taken into account (Endersby *et al.*, 2007). Microsatellite loci in lepidopterans have also been associated with transposable elements causing whole loci to vary in copy number (Zhang, 2004), which motivated the development of exon-primed intron-crossing (EPIC) markers with primers that bind to conserved exon sequences, reducing the frequency of null alleles and allowing the characterization of the more variable intronic sequences (Tay *et al.*, 2008). EPIC markers also have the advantage of applicability across related species.

Here we examine nucleotide diversity and LD at nine Z-linked EPIC markers in *H. armigera*. *H. armigera* follows a ZZ/ZW sex determination with the female being the heterogametic sex. The focus on Z-linked loci meant that we were able to Sanger-sequence amplicons directly from females and thereby (i) prevent insertion/deletion heterozygosity from confounding sequence traces and (ii) measure the extent of LD directly without having to infer gametic phase. This enabled us to quantify the extent of genomic nucleotide diversity within *H. armigera* and determine whether LD declines in *H. armigera* over short distances as it is reported to decline in *B. mori*. In doing this, we identified a locus exhibiting unusual patterns of nucleotide diversity. In order to determine whether a selective sweep was the best model to describe the patterns in this locus, we sequenced two additional flanking loci, sequenced the gene in related species and characterized the global distribution of the putative sweep haplotype.

MATERIALS AND METHODS

To examine baseline nucleotide diversity in *H. armigera*, we chose EPIC markers from the Z chromosome. Loci were chosen according to the following criteria: they were likely to be located on the Z chromosome, likely to be dispersed across the Z chromosome, we were confident in their gene models, their introns were of a size appropriate for reliable PCR amplification, and had flanking exon sequence conservation. We chose to examine *Apt* and *Tpi* because they had been reported to be Z-linked in Lepidoptera species (Jiggins *et al.*, 2005; Yasukochi *et al.*, 2006; d'Alencon *et al.*, 2010). Two P450 genes were chosen but neither were expected to have a role in insecticide resistance as their orthologs have roles in development (Willingham and Keil, 2004; Feyereisen, 2005). The other loci were chosen without regard to gene ontology.

Development of Z-linked EPIC markers

B. mori Z-linked proteins were selected from the silkworm genome database (http://www.silkdb.org/cgi-bin/silkgo/index.pl) and cross-checked against Genbank accessions to obtain more detailed annotations. Protein sequences were checked against the *B. mori* genome to ensure that they were single copy and subsequently used to identify orthologs in the *H. armigera* contig database (*Helicoverpa* Genome Consortium, unpublished) under the *protein2genome* model in Exonerate (Slater and Birney, 2005). Other criteria were to avoid loci that mapped to the ends of contigs (because sequence quality could be compromised) and regions containing repeat sequences.

A total of nine *B. mori* proteins with a BLASTX score of at least 200 were shortlisted so as to include all five scaffolds of the *B. mori* Z chromosome (nscaf1690, nscaf2210, nscaf2734, nscaf3040 and nscaf3068). EPIC markers were designed to span at least one intron with product sizes ranging from 600 to 1200 bp with exon sequences that were at least 50 bp away from either end of an intron. The loci *down3* (downstream 3 kb) and *up3* (upstream 3 kb) were subsequently included after the patterns of variation around *Cyp303a1* were observed although they did not return any matches to known protein-coding sequences.

Twenty families were generated from single-pair matings of a laboratory-maintained colony to follow the markers through a pedigree. F1 individuals were sexed as pupae and re-assessed as adults. This colony was initiated from field samples collected in the vicinity of Toowomba, Australia (27°34′S, 151°57′E) in 2002 but has since been subjected to multiple injections of another laboratory-maintained strain, GR, to counteract the effects of inbreeding depression (Mahon *et al.*, 2008). A separate set of pedigrees was used to ascertain Z-linkage of the *Cyp303a1* alleles as our colony had fixed for the *Ins200* allele. These families were derived from field samples collected in the MacIntyre Valley and obtained as ethanol-preserved moths (parents) and pupae (offspring). DNA extractions were carried out using a standard phenol–chloroform procedure. Scoring was performed by visualizing PCR products on agarose gels.

Samples

A total of 199 *H. armigera* (26 Australian and 173 non-Australian), 5 *H. assulta* and 20 *H. punctigera* DNA samples were obtained from G Behere; data on the collection of these samples is outlined in Behere *et al.* (2007). The Australian *H. armigera* data set consisted of 16 females from Dalmore, Victoria (38°11′S, 145°25′E) and 10 females from Orbost, Victoria (37°42′S, 148°27′E), both of which come from samples from the work by Behere *et al.* (2007) and a new collection of 112 females from MacIntyre Valley, Queensland (28°32′S, 150°18′E). The three collection sites are all temperate agricultural regions. However, they are currently classified into distinct bioregions (of which there are 89 in Australia; *http://www.environment.gov.au/land/nrs/science/ibra*). The Orbost and Dalmore samples were collected off corn and the MacIntyre Valley population off cotton. The Victorian female samples were identified by inferring hemizygosity from Sanger sequencing of Z-linked loci, that is, if overlapping traces (indicating heterozygosity for small indel polymorphisms) were present in the chromatograms, the sample would be designated 'male'. The MacIntyre Valley samples were collected as eggs from the field in 2010 and laboratory reared; sex was determined directly from visual inspection. The Victorian and MacIntyre Valley collections thus represent spatially and temporally separated populations of Australian *H. armigera*. For the genotyping of *Cyp303a1* in non-Australian samples, the data set consisted of 35 individuals from Burkina Faso, 40 individuals from Uganda, 32 individuals from China, 12 individuals from Pakistan and 54 individuals from India. The sexes of these individuals were not known.

Sequencing reactions and quality checks

Cycling conditions varied slightly depending on the targets but were generally 35–40 cycles of 94 °C for 30 s, 60–65 °C for 30 s and 72 °C for 1.5 min. All PCR reactions were carried out using NEB Standard *Taq* polymerase and buffer (catalog number M0273). The final concentration of reagents was 0.025U/μl polymerase, 1x buffer, 200 μM dNTPs, 0.3 μM forward primer and 0.3 μM reverse primer. Sanger-sequencing of PCR products was performed on an ABI3730XL system (Macrogen, Korea).

Sequence quality checks were carried out using Sequencher 4.72 (Gene Codes, Ann Arbor, MI, USA), and sequences were manually edited to match the consensus-by-majority sequence if the base confidence was <40%. This approach was adopted to remove polymorphisms likely to be introduced by sequencing errors, especially for single-nucleotide indels occurring in a homopolymeric run. The disadvantage is that true polymorphisms occurring at low frequency are potentially discarded, but the preference was to adopt a conservative estimate of polymorphism given our expectations of a high-diversity genome.

Sequence diversity and nucleotide divergence

For *H. punctigera* and *H. zea* sequences, a repository was available for BLAST searches and accessing contigs (*Helicoverpa* Genome Consortium, unpublished). In the case of *H. assulta* where no such database could be interrogated, PCR reactions were carried out with *H. armigera* primers and Sanger-sequenced; orthology is assumed because only a single specific product was amplified. Multiple sequence alignment was performed using Seaview 4.0 (Gouy *et al.*, 2010) and ClustalX (Larkin *et al.*, 2007). Interspecies alignments were carried out in a two-step process: first, by defining each intraspecies alignment as a profile and, second, by aligning the profiles using the profile alignment option in ClustalX. Maximum-likelihood trees were constructed using PhyML under a GTR model, with support for clades based on 100 bootstrap replicates. Analyses of polymorphism and LD were carried out using DnaSP 5.10.01 (Librado and Rozas, 2009), with alignment files indicated as haploid Z chromosome. Estimates of polymorphism and divergence presented are uncorrected with respect to models of DNA evolution. Population differentiation was evaluated using an unbiased estimator of *FST* proposed by Hudson *et al.* (1992).

Linkage disequilibrium

LD was estimated as the square of the correlation coefficient, r^2, for each pair of single-nucleotide polymorphisms using only parsimony-informative sites; sites segregating for three or four nucleotides and all indel polymorphisms were ignored. The statistical significance of each pairwise comparison was evaluated using Fisher's exact test and the χ^2 test followed by Bonferroni correction for multiple testing. The number of significant pairwise comparisons as evaluated by Fisher's exact test (Supplementary Table S1) is more conservative, but we wanted to relax these constraints given our hypothesis of a low-LD genome; hence LD heatmaps were plotted using the outcomes of the χ^2 test, which tended to evaluate a higher number of results as significant. The heatmap for *Cyp303a1* and its flanking regions (3 kb upstream and downstream) was plotted by concatenating the sequences of *up3*, *Cyp303a1* and *down3* in individuals (five insertion and six deletion alleles) where data were available for all three loci. Heatmaps were visualized using the LDheatmap package in R (http://www.r-project.org/).

Decay of LD over physical distance was modeled on the expectations of Hill and Weir (1988) and implemented with the nonlinear least-squares function in R.

Coalescent simulations

A Monte Carlo program, msms (Ewing and Hermisson, 2010) was used to generate samples evolving under a neutral infinite-sites model based on the coalescent process, assuming a large and constant population size. All simulations were performed using the sample size n and number of segregating sites S as minimal input parameters. The value of $n=63$ was chosen to reflect the allele frequencies in field populations while maximizing the number of sequenced alleles in the analyses, that is, 44 *Ins200* and 19 *Del200* alleles (defined below). The value of $S=80$ was obtained from empirical data (Table 1). The recombination parameter C was estimated using two methods: the number of minimum recombination events (R_m) using the method of Hudson and Kaplan (1985), and R from Hudson (1987), which is based on the variance of the average number of differences between pairs of sequences in a sample.

For the simulations under a single-locus selection model, effective population sizes (N_e) between 10^5 and 10^7 and an allele frequency of 0.3 were used with the SF switch, with time t set to 0 to represent selection occurring up to the present time. The number of data sets in Table 3 (D) were chosen from simulations that resulted in the highest probability (typically $N_e=10^7$) so as to maximize the available data for subsequent analyses. Two values of the selection coefficient, s, were tested: the first representing weak-to-modest levels of selection ($s=0.01$) and the second representing a strong positive selection ($s=0.1$). The effect of the beneficial allele on the heterozygote was set to be half of that of the homozygote, that is, -SAA $2N_es$ -SAa $1N_es$.

RESULTS

Development of Z-linked EPIC markers

The development of nine *H. armigera* Z-linked markers in this study (Supplementary Table S2) was informed by previous reports of synteny in lepidopterans (Jiggins *et al.*, 2005; Yasukochi *et al.*, 2006; d'Alencon *et al.*, 2010). *Apt* and *Tpi* have been established as Z-linked

Table 1 Nucleotide diversity and Tajima's *D* for nine loci surveyed in this study

Locus	n	*No. of sites (bp)*[a]	S[b]		π		*Tajima's* D	
Apt (55)	Dalmore (8)	751–772	26	81	0.01	0.02	0.57	−0.99
	Orbost (5)		17		0.01		1.48	
	M. Valley (41)		80		0.02		−1.05	
Cycle (21)	M. Valley (20)	824		40		0.01		−0.02
Cyp303a1 (83)[c]	Dalmore (14)	470–515	58	80	0.05	0.05	1.39	1.77
	Orbost (10)		50		0.05		2.27*	
	M. Valley (56)		76		0.05		1.85	
Cyp305b1 (22)	M. Valley (21)	649		56		0.02		−0.30
Period (36)	Dalmore (13)	336–511	36	54	0.03	0.02	−0.79	−1.57
	Orbost (5)		29		0.03		−0.71	
	M. Valley (17)		52		0.03		−1.67	
Phc (36)	Dalmore (12)	482–534	64	94	0.04	0.04	−0.47	−0.48
	Orbost (7)		61		0.04		−0.39	
	M. Valley (16)		77		0.04		−0.19	
SCAP (12)	M. Valley (11)	840		104		0.03		−0.79
Tc (16)	M. Valley (15)	817		60		0.02		0.43
Tpi (33)	Dalmore (11)	514–544	100	125	0.07	0.06	0.08	0.04
	Orbost (2)		35		0.06		NA	
	M. Valley (19)		115		0.06		−0.24	

Where estimates are presented in two columns under a single heading, the left column represents estimates for an individual population while the right column represents estimates after pooling sequences of all three populations. Figures in brackets after the locus name represent the total number of sequences surveyed, including the reference strain. Tajima's *D* for the Orbost population of *Tpi* is not available as a minimum of four sequences are required. *$P<0.05$.

[a]The number of sites is presented as a range due to the differing subsets of indel polymorphisms present in different populations. As gapped sites are excluded from this analysis, the lower boundary represents the number of sites considered when alleles from all three populations are pooled.

[b]Number of segregating sites, including singletons.

[c]Includes two sequences from a laboratory-maintained colony.

loci in multiple species, whereas the other loci were chosen because they were single-copy sequences that had 1:1 orthologs on *B. mori* Z-linked genes. Sex-limited inheritance of PCR amplicon size variation across pedigrees confirmed Z-linkage for *Cyp303a1*, *Phc* and *Period* (Supplementary Figures S1 and S2). For the remaining four loci, direct sequencing of amplicons was carried out on female samples without pedigree analyses. The absence of overlapping traces in the chromatograms indicated that the sequences were hemizygous and confirmed that these four loci were also on the Z chromosome.

Sequence diversity

We initially characterized five loci for which Z-linkage was determined by pedigree analyses (*Apt*, *Cyp303a1*, *Period*, *Phc* and *Tpi*) in two Victorian population samples described by Behere *et al.* (2007). Consistent with mitochondrial DNA analyses by Behere *et al.* (2007), we found no evidence for structure between these two populations ($FST<0.06$ at all loci examined). These initial results prompted us to expand the data set by obtaining an additional Australian population. To avoid the redundancy of work associated when scoring males (see Materials and methods section), we obtained 112 adult females from McIntyre Valley (on the border of New South Wales and Queensland) and scored them at all nine Z-linked loci. There was no evidence for population structure between the McIntyre Valley samples and the two Victorian populations (Supplementary Table S3), which suggested that LD analyses could be conducted on alleles pooled from all three populations (see below).

Levels of nucleotide diversity across all loci and the three Australian collection sites were high (694 single-nucleotide polymorphisms in <6 kb of sequence) and did not differ substantially between collection sites for any locus (Table 1). However, π values differed up to sixfold across loci (0.01–0.06 nucleotide differences per site) while indel variation differed by up to sevenfold across loci (0.002–0.014 indel events per site; Supplementary Table S4). Haplotype diversities were in the range of 0.7–1 for each locus per location, and we did not observe significant geographic structuring of haplotypes. Six of the nine loci had a negative Tajima's *D*, indicating an elevated number of rare variants in the samples although statistically the values were non-significant (consistent with the neutral model). The most notable feature of the frequency spectrum analysis was that *Cyp303a1* had a highly positive Tajima's *D*. This was also true when each population was looked at individually, although only the Orbost population crossed the standard significance threshold.

Linkage disequilibrium

LD was calculated after pooling alleles from all three populations to maximize the sample sizes and thereby increase the power to detect significant associations. The level of LD in *H. armigera* was generally very low and of a similar magnitude to that seen in *B. mori*. LD was found to halve within 200 bps at each locus with the exception of *Cyp303a1*, whereby the distance at which r^2 reached half its maximal estimated value was beyond the size of the 600/800 bp sequenced region (Figure 1). The paucity of LD at *Phc* and *Tpi* is striking given the total number of comparisons involved (Figure 2, Supplementary Table S1) as there are 63 and 77 parsimony-informative sites in *Phc* and *Tpi*, respectively.

Signals of selection at *Cyp303a1*?

The positive Tajima's *D* values and the excessive LD at *Cyp303a1* can be further understood by the allelic network of this locus relative to that of the other loci (Figure 3). An unrooted maximum-likelihood tree reveals an anomalous long branch separating two *Cyp303a1* haplogroups that we will refer to as *Del200* and *Ins200* because a diagnostic feature of the two haplogroups is a 200-bp indel. The other surveyed loci exhibit more gradations in their phylogenies. *Tpi* does have a long internal branch but that can be attributed to a single stretch of 25 nucleotides containing five fixed differences. Even after exclusion of the 200-bp indel in *Cyp303a1*, the long internal branch is still apparent (Supplementary Figure S3) as there are 31 other fixed differences that are interspersed throughout the sequence alignment. The extent of divergence between the *Ins200* and *Del200* haplogroups raised concerns as to whether they represented two allelic types or paralogs. However, we confirmed that the indel polymorphism segregated in an allelic Z-linked manner where female offspring always presented only one copy of the locus, inherited from the male parent (Supplementary Figure S2).

A total of 44 *Ins200* and 39 *Del200* alleles were sequenced so that the level of variation within haplogroups could be compared with the divergence between haplogroups (Table 2). The *Del200* haplogroup contained very short branch lengths and was dominated by a single haplotype (32 of the 39 individuals). Within the *Del200* haplogroup, there were very low levels of variation, evident in the small number of segregating sites and haplotypes compared with an equivalent number of *Ins200* alleles. No indel polymorphisms were observed within the *Del200* haplogroup and Tajima's *D* was significantly negative. In contrast, the *Ins200* haplogroup had levels of nucleotide and indel diversity of a similar magnitude to that of other loci. The π value of the *Ins200* haplogroup was 20 times that of the *Del200* haplogroup, and multiple smaller indels were present within the *Ins200* haplogroup. The divergence between haplogroups exceeded levels of nucleotide diversity at all other loci examined in this study.

To assess whether evolution at the *Cyp303a1* locus was compatible with the neutral model, two tests were carried out using coalescent simulations. The first test addressed the likelihood of obtaining i identical alleles from a sample of size n given the diversity of the sample (Hudson *et al.*, 1994). By analyzing the length of the amplicons, the frequency of the *Del200* haplogroup was observed to be approximately 28% in all three populations. The preceding haplotypic analysis indicated that 32/39 or 82% of alleles would fall under a single haplotype, hence the value of i was determined to be 14 ($0.28\times0.82\times63$) for a sample size of $n=63$. Simulated data sets were generated under three scenarios: (i) no recombination, (ii) $C=R_m$, the minimum number of recombination events which underestimates the total number of recombination events (Hudson and Kaplan, 1985), and (iii) $C=R$ estimated from the method of Hudson (1987). The recombination parameter, C, was estimated using the *Ins200* haplogroup data. The first two scenarios are conservative, yet the probability of obtaining a subset of identical alleles does not exceed 12%. Under the third scenario, which includes a modest amount of recombination, the probability is significant enough to reject the neutral model ($P<0.013$; Table 3 (A)).

The second test addressed the likelihood of observing a major haplogroup that is highly divergent from all other alleles in the population. The test was only conducted on data sets that fulfilled the criteria of the first test ($i\geqslant14$) and was implemented as follows: the pairwise distances (d) between the major haplotype and all others in the data set were calculated. Haplotypes that diverged at <5 substitutions represent variants in one haplogroup (corresponding to the *Del200* haplogroup) while those containing >30 substitutions (fixed differences) represent haplotypes from other haplogroups (akin to the *Ins200* haplogroup). Data sets containing any values of $5<d<30$ are considered unlike our observed data set, that is, are not divided into diverged haplogroups. The simulations conducted in

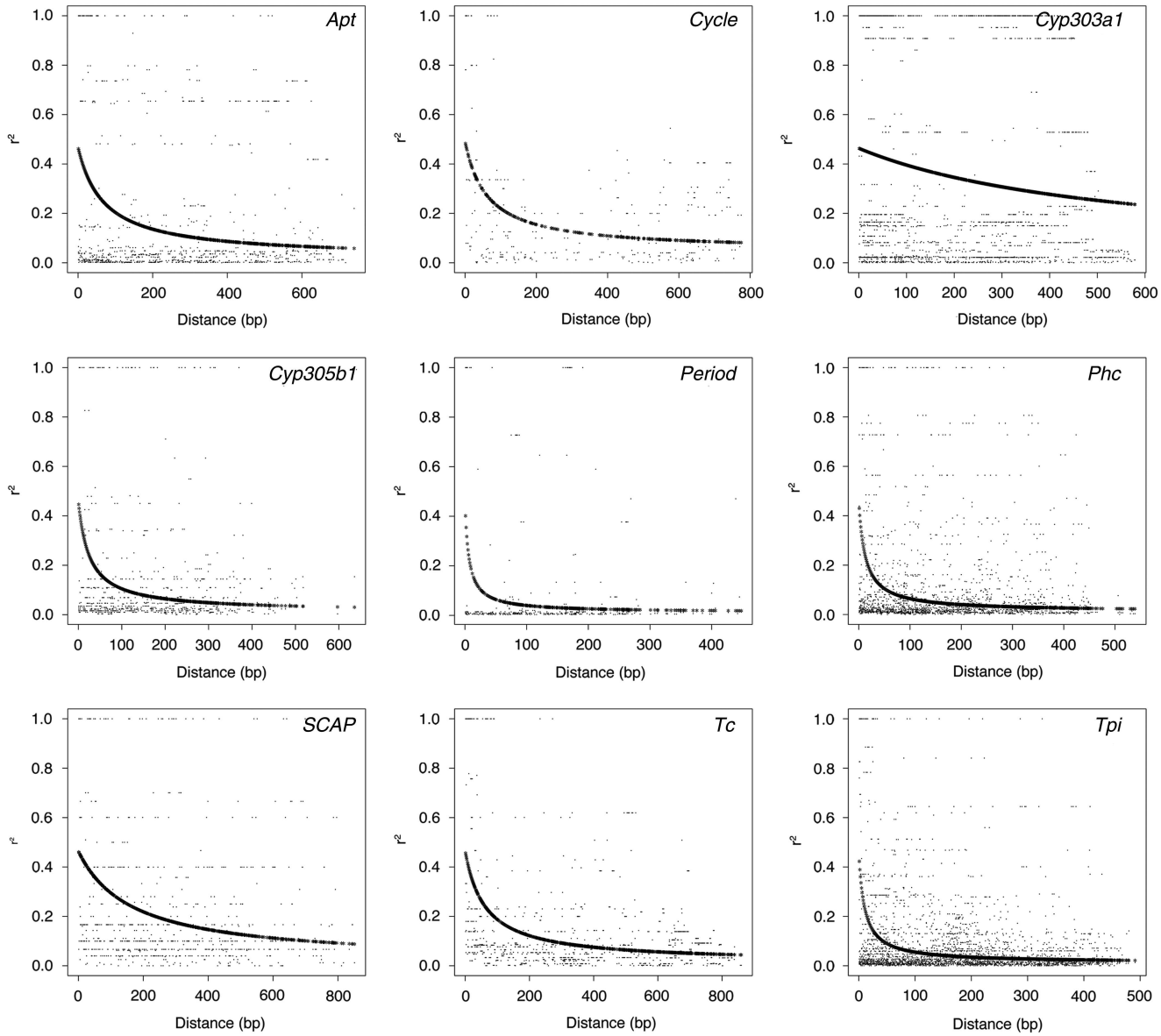


Figure 1 Plots of r^2 over physical distance in base pairs, with curves showing the decay of LD modeled on the expectations of Hill and Weir (1988). The approximate distance at which $E(r^2)$ decays to 0.2 (approximately half the maximum estimated value across all nine loci) for each decay curve is 104, 122, >579, 25, 7, 14, 186, 76 and 11 bp for *Apt*, *Cycle*, *Cyp303a1*, *Cyp305b1*, *Period*, *Phc*, *SCAP*, *Tc* and *Tpi*, respectively. Maximum values of $E(r^2)$ ranged from 0.39 to 0.51. The number of alleles sampled for each locus is reported in Table 1.

the absence of recombination recovered diverged haplogroups (like the observed data set) 10% of the time, whereas those with even the minimal level of recombination recovered the diverged haplogroups <1% of the time (Table 3 (B)). These simulations demonstrate that, under a strict neutral model of coalescence, it is highly unlikely to observe a haplotype network divided into divergent haplogroups with one containing limited diversity.

To assess whether selection was the most parsimonious explanation for both the low diversity of the *Del200* haplogroup and the divergence between the two haplogroups, the two tests outlined above were carried out under two models incorporating the selection coefficient parameter, s (Table 3 (C–F)). Using an effective population size (N_e) between 10^5 and 10^7, allele frequency (f) of 0.3 and allowing selection to occur to the present time ($t=0$), the likelihood of obtaining 14 identical alleles out of 63 was >85% in all the scenarios tested (Table 3 (C and E)). However, the likelihood of obtaining a long internal branch fell to <1% when even a minimal amount of recombination was allowed, irrespective of the strength of selection (Table 3 (D and F)). In the absence of recombination, the diverged haplogroups were recovered approximately 5% of the time. These simulations suggest that positive selection in and of itself is insufficient to account for the patterns observed in our data set, and some secondary mechanism affecting recombination may have accompanied the selective event.

To further investigate whether selection affected *Cyp303a1* and the *Del200* haplogroup in particular, the frequency of the deletion allele outside of Australia was examined using PCR amplicon length analysis. The deletion haplogroup was not detected in any samples from India ($n=54$), Pakistan ($n=12$), Burkino Faso ($n=35$) or Uganda ($n=40$). However, three deletion alleles were present in the

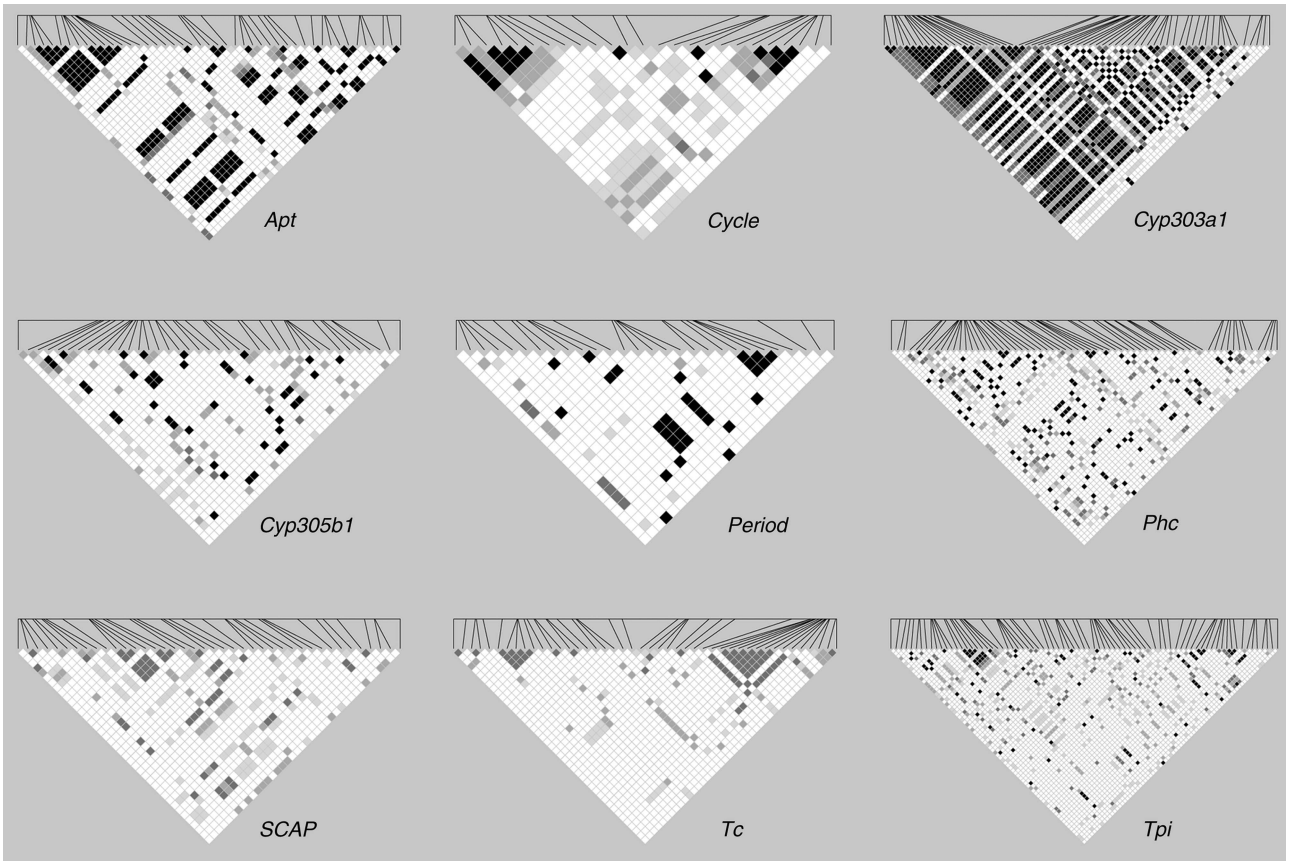


Figure 2 LD heatmaps for each of the nine loci plotted as significance of the r^2 value in pairwise comparisons of segregating sites. Only bi-allelic sites are included. In all cases, the physical distance between the first and last site does not exceed 1 kb. Shading indicates significance level with black: $P<0.001$ (significant by Bonferroni), 80% grey: $P<0.001$, 50% grey: $0.001<P<0.01$, 20% grey: $0.01<P<0.05$, white: not significant.

Chinese population ($n=32$) of which one was confirmed by Sanger sequencing. This Chinese *Del200* allele differed from the major *Del200* haplotype found in Australia at a single site. A low frequency of novel amplicon lengths were found in these global samples but most were approximately 800 bps, which is typical of the *Ins200* haplogroup. Thus this locus contrasts dramatically with the multiple markers from multiple studies that exhibit low *FST* in global populations (Daly and Gregg, 1985; Nibouche *et al.*, 1998; Zhou *et al.*, 2000; Behere *et al.*, 2007).

We postulated that the target of selection resulting in the patterns observed in the *Del200* haplogroup may not be within the sequence of the EPIC amplicon we surveyed but another polymorphism that is in LD with it. To determine whether there were any amino-acid changes at *Cyp303a1* that could be the variant targeted by selection, the complete coding region was sequenced from two *Ins200* and two *Del200* alleles of a laboratory-maintained colony. We did not detect any non-synonymous substitutions that could discriminate the two alleles, although there were four synonymous polymorphisms segregating.

To determine whether the polymorphism targeted by selection could be limited to the *Cyp303a1* locus, adjacent regions 3 kb upstream and downstream of *Cyp303a1* were sequenced from a subset of the field samples (5:7 and 11:12 *Ins200*:*Del200* alleles, respectively). An LD heatmap illustrates that approximately 300 bp of the 3′ end of the upstream region is in LD with the indel (Supplementary Figure S4). Levels of diversity at this locus were similar to that of other loci with $\pi=0.05$ and $\pi(i)=0.010$ and Tajima's D was slightly negative (−0.29) but not statistically significant. In contrast, the downstream region (*down3*) exhibited very low diversity ($\pi=0.004$) with no indels and had a significantly negative Tajima's D (−1.95, $P<0.05$). Thus selection may indeed be acting downstream of *Cyp303a1*.

Divergence from other species

The lack of nucleotide polymorphism in the downstream locus is consistent with the idea that there has been a selective sweep in the vicinity, and the implication is that this downstream locus is closer to the target of selection. However, an alternate hypothesis is that strong purifying selection independent of the effect seen at *Cyp303a1* acts upon this downstream locus, even though it appears to be non-coding (that is, sequences are constrained because any change alters an important function). The divergence between closely related species could help us discriminate between these two hypotheses. If such constraint is acting on the sequence 3 kb downstream of *Cyp303a1*, then it may have been acting since the divergence of the species. Table 4 shows the divergences between species for the nine loci and the region downstream of *Cyp303a1*. As expected from the species phylogeny, *H. armigera* sequences showed the greatest divergence with *H. punctigera* followed by *H. assulta* and then *H. zea*. *H. zea* appeared very similar to *H. armigera*. The *down3* locus exhibited values that were similar to those of other loci, which rejects the hypothesis that the lack of diversity observed in *H. armigera* could reflect excessive purifying selection at that sequence over the period encompassing the divergence of these species. Rather, the lack of diversity despite

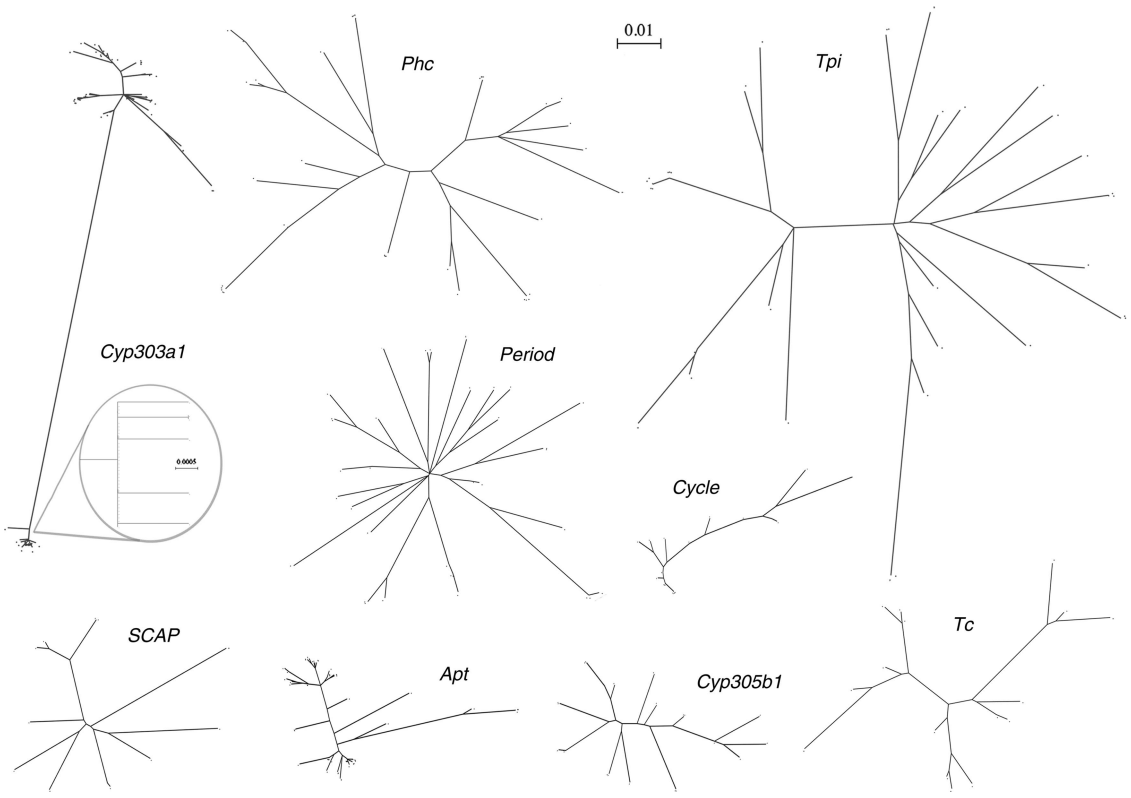


Figure 3 Unrooted maximum-likelihood trees for each of the nine loci. Each tip represents an allele. A long branch separates the two clades representing the *Ins200* (above) and *Del200* (below) alleles at *Cyp303a1*. Inset: The clade containing the *Del200* alleles has very short branch lengths, with 32 out of 39 individuals carrying the same haplotype.

Table 2 Comparison of the *Cyp303a1* Del200 and Ins200 haplogroups

Statistic	*Del200*	*Ins200*
Number of sequences, *n*	39	44
Number of sites, excluding gaps	523	681
Segregating sites (including singletons), *S*	9	63
Parsimony informative sites	1	45
Number of haplotypes, excluding gaps	7	32
Number of indel events, *I*	0	17
Nucleotide diversity per site, π	0.001	0.019
Tajima's *D*	−2.23**	−0.43
Between haplogroups		
Number of fixed differences	31	
Nucleotide divergence	0.09	

**$P<0.01$.

divergence favours a model of recent positive selection affecting loci in this region.

Furthermore, rather than showing selective constraint, the *Cyp303a1* amplicon shows high divergence relative to the values observed at other loci in *H. assulta* and *H. zea*. The divergence across the intron suggests overlapping indels during the evolution of this locus, making alignments difficult. For instance, the *H. punctigera* sequence shared one characteristic of the deletion variant in that it was lacking the 200-bp insert. However, it also contained a 100-bp deletion in a region common to both subgroups and various smaller indel polymorphisms in the regions flanking the *H. armigera* 200-bp indel. The *H. assulta* sequence had an 800-bp insertion that incorporated the 200-bp *H. armigera* insertion. The *H. zea* sequence incorporated parts of the *H. armigera* insertion. A maximum-likelihood tree of the sequenced region across *H. armigera*, *H. assulta*, *H. punctigera* and *H. zea* (Figure 4) shows low bootstrap support of the relationships between species, except for the placement of *H. punctigera* as the outgroup. We also scored the state of the *H. punctigera*, *H. assulta* and *H. zea* sequences where the fixed differences between the *Ins200* and *Del200* haplogroups occurred (Supplementary Table S5). Although the three outgroup sequences superficially have more states in common with the insertion variant, there is no clear indication that the insertion is the ancestral state.

DISCUSSION

The levels of diversity we observe on the Z chromosome of *H. armigera* (average $\pi = 0.03$) are high relative to genome-wide estimates of diversity in other insects, which are generally high among that of other taxa (Leffler *et al.*, 2012). In *Aedes aegypti*, *Anopheles funestus* and *Anopheles gambiae* for instance, nucleotide diversity in noncoding regions is approximately 0.01; in *Drosophila melanogaster* it is ~0.01 and in *D. simulans* it is ~0.02 (Morlais and Severson, 2003; Wondji *et al.*, 2007; Langley *et al.*, 2012; O'Loughlin *et al.*, 2014). It is worth noting that, if the mutation rates differ between the sexes, then nucleotide diversity at neutral loci can differ between sex chromosomes and autosomes (Vicoso and Charlesworth, 2006). In Lepidopterans where males are the homogametic sex, a Z chromosome spends 2/3 of its evolutionary time in males, whereas an autosome only spends half of its time in this potentially more

Table 3 Coalescent simulations with and without selection

	Recombination parameter		
	C = *0*	C = R_m	C = R
No selection			
(A) Number of data sets	10 000	10 000	10 000
P (*i* ⩾ 14)	0.120	0.064	0.013
(B) Number of data sets	1200	643	126
P (*d* < 5 ∪ *d* > 30)	0.10	<0.01	<0.01
Selection coefficient, s = 0.01			
(C) Number of data sets	1000	1000	1000
P (*i* ⩾ 14)	>0.90	>0.90	>0.85
(D) Number of data sets	976	974	954
P (*d* < 5 ∪ *d* > 30)	0.05	<0.01	<0.001
Selection coefficient, s = 0.1			
(E) Number of data sets	1000	1000	1000
P (*i* ⩾ 14)	>0.95	>0.95	>0.95
(F) Number of data sets	977	968	962
P (*d* < 5 ∪ *d* > 30)	0.04	<0.01	<0.001

Probability of observing (A, C, E) a minimum of *i* identical sequences using a threshold value determined from empirical observations of the frequency of the *Del200* haplogroup and (B, D, F) a major haplogroup that is highly divergent from all other alleles in the population whereby the pairwise distance, *d*, between the major allele and all other sequences in a data set is either <5 (representing variants within the *Del200* haplogroup) or >30 (representing the 31 fixed differences between the *Ins200* and *Del200* haplogroups). All data sets were simulated using the parameters $n = 63$ and $S = 80$. For simulations with selection, N_e ranged from 10^5 to $10^7$and the *SF* option with $t = 0$ and $f = 0.3$ was used (Ewing and Hermisson, 2010). In addition to a no recombination scenario, two estimates of the recombination parameter, *C*, were included: the minimum number of recombination events, $R_m = 5$ (Hudson, 1987), and the estimator based on the variance of the average number of differences between pairs of sequences, $R = 21.7$ (Hudson and Kaplan, 1985).

Table 4 Nucleotide divergence between *H. armigera* and *H. assulta*, *H. punctigera* and *H. zea*

Locus	*Nucleotide divergence, D_{xy}, between* H. armigera *and*		
	H. assulta	H. punctigera	H. zea
Apt	0.06	0.08	0.04
Cycle	0.06	0.06	0.03
Cyp303a1	0.11	0.18	0.12
Cyp305b1	0.04	0.10	0.04
down3	0.07	0.13	0.05
Period	0.05	0.08	0.03
Phc	0.07	0.11	0.03
SCAP	0.10	0.19	0.04
Tc	0.09	0.12	0.05
Tpi	0.07	0.11	0.07

mutagenic sex. Furthermore, positive selection could cause the Z chromosome to evolve faster as recessive alleles are exposed in females (Vicoso and Charlesworth, 2006). Therefore, it is possible that the nucleotide diversity we observe on the Z of *H. armigera* may be elevated relative to the genome-wide value. If estimates from silk moths are used as a guide (Sackton *et al.* 2014), the autosomal diversity will be approximately 60% of that Z chromosomes (that is, ~0.02), and that would not alter our conclusion that *H. armigera* exhibits high levels of nucleotide diversity.

The frequency of insertions and deletions is also higher in *H. armigera* ($\pi(i) = 0.005$) relative to *D. melanogaster* where $\pi(i)$ is <0.003 for intergenic and intronic regions (Ometto *et al.*, 2005). A pragmatic consequence of such a high indel frequency is that direct sequencing of EPIC PCRs in this species may be problematic at autosomal loci (and Z loci in males) because the sequence trace at each frequently spaced indel (every 200 bps) will feature two overlapping sequences (observed as double peaks on the sequence chromatograms) that may be hard to disentangle. At a theoretical level, we are left with the question of whether the high nucleotide diversity ($\pi \approx \theta$) in this species is due to a large effective population size or a high mutation rate ($\theta = 4N_e\mu$).

This study also reveals that the *H. armigera* genome displays remarkably limited LD. For eight of the nine loci characterized herein, r^2 drops to half its estimated maximal value within 200 bps, and this is low relative to that of the other lepidopterans so far characterized and that of other insects (Supplementary Table S6). Regardless of the reasons for the different levels of LD in different species, it creates an important design consideration for future population genomic studies in these insects. For instance, a rigorous genome-wide association study in *H. armigera* would need such a high marker density that whole-genome sequencing might be preferable to technologies that genotype 'tag' single-nucleotide polymorphisms. The high levels of nucleotide diversity coupled with rapid decay of LD also mean that genotype imputation approaches will be limited. Another challenge of allele-rich architecture is genome assembly itself because alleles may be confused as paralogs. However, an advantage of a low-LD, high-diversity genome should be easier identification of causal variants in genome-wide association studies or selective sweep studies.

This study supports previous findings of little population subdivision in Australian *H. armigera* (Daly and Gregg, 1985; Endersby *et al.*, 2007). Low *FST* values at multiple loci sampled from spatially and temporally different populations suggest extensive gene flow. The paucity of LD is consistent with this scenario—if the three populations were genetically differentiated, we would expect a modest degree of significant associations due to 'admixture' from the pooling of alleles. The presence of a single haplotype that appears to have recently arisen to similar intermediate frequencies (the *Del200* haplotype) in geographically separated samples is parsimoniously explained by extensive gene flow in Australia.

The divergence data reported here also suggests that *H. zea* are not substantially diverged from *H. armigera*. For instance, divergence between *H. armigera* and *H. zea* at the *Phc* locus (0.03) was less than that observed between some *H. armigera* alleles ($\pi = 0.04$). This is consistent with the origin of *H. zea* from within an ancestral *H. armigera* population as proposed by Mallet *et al.* (1993) and affirmed by Behere *et al.* (2007).

A footprint of a selective sweep?

The *Cyp303a1* locus exhibits multiple patterns that are aberrant relative to the other loci surveyed here and that are inconsistent with neutral expectations. Among these is extended LD and an unusual frequency spectrum of polymorphisms. These patterns can be attributed to the occurrence of two divergent haplogroups, one of which seems to have recently arisen to high frequency in Australian populations as it exhibits very little allelic diversity despite it being at 28% frequency. The coalescent simulations performed here show that such patterns are extremely unlikely in a neutral model, particularly when so much recombination is observed in the *H. armigera* genome. As discussed below, in order to see such patterns, the extent of recombination among the sampled alleles must have been distorted by the influence of selection, a molecular mechanism limiting the site of recombination at meiosis, and/or population demographics.

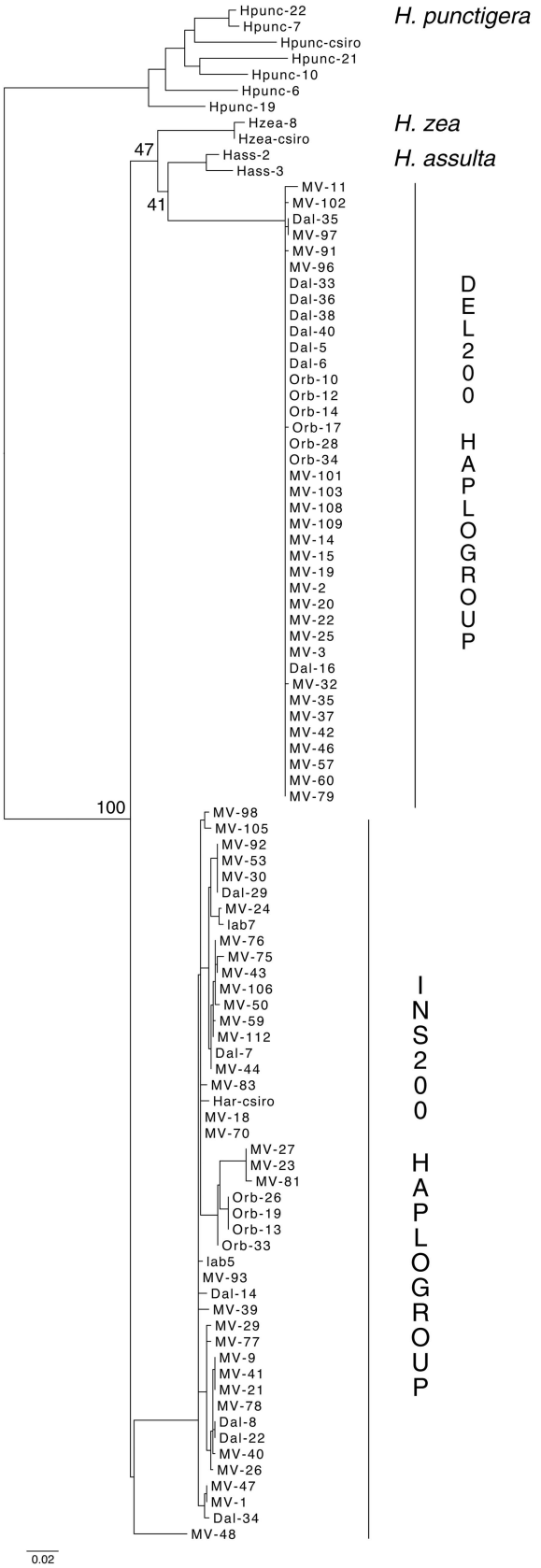


Figure 4 Maximum-likelihood tree of *Cyp303a1* sequenced region from *H. armigera*, *H. assulta*, *H. punctigera* and *H. zea*. For clarity, only bootstrap values pertaining to the relationships between species are shown.

A subset of highly similar alleles within a set of diverged alleles has been reported at particular loci in other species and is often attributed to partial selective sweeps—the model is that a favourable variant has increased in frequency at such a rate that recombination has not had time to occur, enabling nearby variants to 'hitchhike' to intermediate frequency (Hudson *et al.*, 1994; Schlenke and Begun, 2004; Sanchez-Gracia and Rozas, 2007). In a similar manner, the lack of variation in our deletion haplotype is inconsistent with a scenario of neutral polymorphism at intermediate frequency or the scenario of an old polymorphism maintained through balancing selection. Instead, the selective sweep model is supported by the *Del200* haplogroup displaying a skewed frequency spectrum (Tajima's *D* is significantly negative), which can be interpreted as recent positive selection resulting in an intermediate frequency of the polymorphism. Furthermore, the pattern of polymorphism in a noncoding region 3 kb downstream of this locus also shows a significantly negative Tajima's *D* value. However, there is no clear bifurcation into haplogroups at this downstream locus. This could be explained by a hard sweep focussed on a variant closer to the *down3* locus, purging variation at it; the *Cyp303a1* intron that is further away from this selective pressure thus sits on the 'shoulder' of the sweep where a limited amount of recombination has prevented fixation of the *Del200* haplogroup.

A second explanation for the patterns observed at *Cyp303a1* is that recombination is suppressed in the *Del200* haplogroup because of an uncharacterized molecular feature such as an inversion or perhaps the 200-bp indel itself has prevented exchange between chromosomes at prophase I of meiosis. This would explain the second extraordinary feature of the *Cyp303a1* genealogy—the accumulation of so many divergent sites between the *Ins200* and the *Del200* haplogroups (31 fixed differences). Recombination would be unimpeded among the *Ins200* alleles but they could not recombine with alleles from the *Del200* haplogroup. However, such recombination suppression would not explain the Tajima's *D* test results among the *Del200* alleles or the *down3* locus. The molecular explanation for the lack of recombination would therefore still need to be accompanied by a secondary selection event.

A third way that recombination could be distorted is if our samples were influenced by demographic events such that alleles were not sampled from a population where random mating had been occurring throughout the history of their coalescence. We have already noted that most of the data presented here are consistent with previous suggestions of little population structure in *H. armigera*. However, the population structure at the *Cyp303a1* locus is exceptional in that the *Del200* haplogroup is present at high frequency in all Australian samples yet does not occur in African, Indian or Pakistani populations and is at very low frequency in the Chinese population we surveyed. These data support the model that the *Del200* haplogroup arose in Australia and has increased to its current frequency of 28% due to positive selection and has spread to China. The alternate model, separating the originating country (for example, China) from the sweep to high frequency, implies that the selective agent driving the sweep is geographically limited to Australia; this is a more complex and therefore less likely scenario.

We have not surveyed the other Z-linked loci outside Australia, yet mitochondrial DNA, allozyme, microsatellite and EPIC PCR analyses do not suggest that Australia houses particularly divergent or isolated alleles (Daly and Gregg, 1985; Nibouche *et al.*, 1998; Behere *et al.*, 2007; Endersby *et al.*, 2007; Tay et al., 2008). So if there is no evidence of a reservoir of diverged alleles in Australian *H. armigera*, where does the *Del200* haplogroup come from? One possibility is that it was introduced via introgression from a related species. This notion is

appealing because it explains the high level of divergence between the two haplogroups. The term 'comet allele' has been proposed to describe haplotypes that have introgressed across species or sub-species boundaries—similar to comets, they have 'dipped' into this system from another lineage (Staubach *et al.*, 2012). There are precedents for such events in other species such as that described by Brand *et al.* (2013) where *D. simulans* alleles have entered the *D. sechellia* genome in an adaptive process. The divergence between the two haplogroups is as great as between *H. armigera* and *H. assulta* at other loci. Given the observed frequencies of the *Del200* haplotype in our Asian and African populations, an Australian origin appears most likely and is consistent with the hypothesized radiation of heliothines on this continent (Matthews, 1999). Our data do not provide evidence for a source population from *H. assulta*, *H. punctigera* or *H. zea*. However, population structures and levels of diversity in *H. punctigera* and *H. assulta* are not as well characterized —the provenance of the *H. armigera* divergent allele could be a cryptic race or isolated population of either species given their overlapping ranges. Alternatively, it could be from another species not characterized here such as *Helicoverpa hardwicki*, *Helicoverpa prepodes* and members of the genus *Australothis* and *Heliocheilus*, which are endemic to Australia (Matthews, 1999; Cho *et al.*, 2008).

Thus to explain the patterns we see at the *Cyp303a1* locus, we are left with two alternate hypotheses both of which involve a selective sweep. In the first, our coalescent simulations suggest that the high level of diversity within *H. armigera* coupled with a molecular-based suppression of recombination in the *Del200* haplogroup may have allowed the emergence of a highly divergent and recently adaptive allele. Alternatively, an adaptive introgression of this locus from another species would explain the sweep of a highly diverged allele through Australian populations.

Finally, we note that the selective agent believed to be driving the patterns in the genealogy of *Cyp303a1* is unknown. Given that *Cyp303a1* is a cytochrome P450 gene, insecticides could be candidates because genes in the P450 multigene family are frequently associated with insecticide resistance (Feyereisen, 2005). However, *Cyp303a1* is a strict (1:1) ortholog to a *Drosophila* gene that has been functionally characterized as being essential for mechanosensation and chemosensation and is expressed only in the sensory bristles (Willingham and Keil, 2004). The occurrence of 1:1 orthology across this taxonomic distance is notable given the multiple gene gain and loss events commonly observed in multigene families and supports the grouping of *Cyp303a1* with the developmental rather than detoxification class of P450s. If *Cyp303a1* is the target of selection, the causal variant would have to be a regulatory mutation as there are no amino-acid differences between the *Ins200* and *Del200* haplotypes, and there was no copy number variation detected at this locus. If insecticide selection is acting on the function of the *Cyp303a1* locus, then a sensing function may be more likely than a detoxifying one. The second possibility is that the target of selection is another gene and *Cyp303a1* is merely a 'hitchhiker', although analysis of the contig on which *Cyp303a1* is located has not revealed any genes in the region extending 10 kb downstream (unpublished data).

In conclusion, this study established that the *H. armigera* genome exhibits high levels of nucleotide diversity within populations and generally high levels of recombination and gene flow yet we discovered an instance where deviations from these trends suggest that footprints of selection can be detected. Genome-wide scans for signals of selection are a complementary approach to genome-wide association study in identifying candidate genes for phenotypes of interest. Evaluating the role of demographic processes in shaping genome architecture remains a major challenge, and new tests for identifying selection will need to accommodate more complex scenarios potentially including introgression from other species.

DATA ARCHIVING

The sequences for Cyp303a1 have been submitted to GenBank (accession numbers KR709083-5). All other sequences are available in the Dryad repository under the doi:10.5061/dryad.123qg.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the leaders of the *Helicoverpa* Genome Consortium for permission to use the data ahead of publication, Ganesh Behere and Wee Tek Tay for providing access to their frozen collections and David Clarke and Robert Good for bioinformatic assistance. SVS was supported by a University of Melbourne Research Scholarship and the CSIRO.

Behere G, Tay W, Russell D, Heckel D, Appleton B, Kranthi K *et al.* (2007). Mitochondrial DNA analysis of field populations of *Helicoverpa armigera* (Lepidoptera: Noctuidae) and of its relationship to *H. zea*. *BMC Evol Biol* **7**: 117.

Brand CL, Kingan SB, Wu L, Garrigan D (2013). A selective sweep across species boundaries in Drosophila. *Mol Biol Evol* **30**: 2177–2186.

Cho S, Mitchell A, Mitter C, Regier J, Matthews M, Robertson R (2008). Molecular phylogenetics of heliothine moths (Lepidoptera: Noctuidae: Heliothinae), with comments on the evolution of host range and pest status. *Syst Entomol* **33**: 581–594.

d'Alencon E, Sezutsu H, Legeai F, Permal E, Bernard-Samain S, Gimenez S *et al.* (2010). Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proc Natl Acad Sci* **107**: 7680–7685.

Daly JC, Gregg P (1985). Genetic variation in Heliothis in Australia: species identification and gene flow in the two pest species *H. armigera* (Hübner) and *H. punctigera* Wallengren (Lepidoptera: Noctuidae). *Bull Entomol Res* **75**: 169–184.

Endersby NM, Hoffmann AA, McKechnie SW, Weeks AR (2007). Is there genetic structure in populations of *Helicoverpa armigera* from Australia? *Entomol Exp Appl* **122**: 253–263.

Ewing G, Hermisson J (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**: 2064–2065.

Feyereisen R (2005). Insect cytochrome P450. *Compr Mol Insect Sci* **4**: 1–77.

Gouy M, Guindon S, Gascuel O (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**: 221–224.

Hardwick DF (1965). The corn earworm complex. *Memoirs Entomol Soc Can* **97**: 5–247.

Harris C, Rousset F, Morlais I, Fontenille D, Cohuet A (2010). Low linkage disequilibrium in wild *Anopheles gambiae s.l.* populations. *BMC Genet* **11**: 81.

Hill WG, Weir BS (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol* **33**: 54–78.

Hudson RR (1987). Estimating the recombination parameter of a finite population model without selection. *Genet Res* **50**: 245–250.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994). Evidence for positive selection in the superoxide dismutase *(Sod)* region of *Drosophila melanogaster*. *Genetics* **136**: 1329–1340.

Hudson RR, Kaplan NL (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

Hudson RR, Slatkin M, Maddison WP (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.

Jiggins CD, Mavarez J, Beltran M, McMillan WO, Johnston JS, Bermingham E (2005). A genetic linkage map of the mimetic butterfly *Heliconius melpomene*. *Genetics* **171**: 557–570.

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE *et al.* (2012). Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* **192**: 533–598.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Laster ML, Hardee DD (1995). Intermating compatibility between North American *Helicoverpa zea* and *Heliothis armigera* (Lepidoptera: Noctuidae) from Russia. *J Econ Entomol* **88**: 77–80.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A *et al.* (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* **10**: e1001388.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Mahon RJ, Olsen KM, Downes S (2008). Isolations of Cry2Ab resistance in Australian populations of *Helicoverpa armigera* (Lepidoptera: Noctuidae) are allelic. *J Econ Entomol* **101**: 909–914.

Mallet J, Korman A, Heckel DG, King P (1993). Biochemical genetics of *Heliothis* and *Helicoverpa* (Lepidoptera: Noctuidae) and evidence for a founder event in *Helicoverpa zea*. *Ann Entomol Soc Am* **86**: 189–197.

Matthews M (1999). *Heliothine Moths of Australia: A Guide to Pest Bollworms and Related Noctuid Groups. Monographs on Australian Lepidoptera,* vol. 7. CSIRO Publishing: Melbourne, Australia. ISBN 0643063056.

Mitter C, Poole RW, Matthews M (1993). Biosystematics of the heliothinae (lepidoptera: noctuidae). *Ann Rev Entomol* **38**: 207–225.

Morlais I, Severson DW (2003). Intraspecific DNA variation in nuclear genes of the mosquito *Aedes aegypti*. *Insect Mol Biol* **12**: 631–639.

Nibouche S, Bues R, Toubon JF, Poitout S (1998). Allozyme polymorphism in the cotton bollworm *Helicoverpa armigera* (Lepidoptera: Noctuidae): comparison of African and European populations. *Heredity* **80**: 438–445.

Nielsen R (2005). Molecular signatures of natural selection. *Annu Rev Genet* **39**: 197–218.

O'Loughlin SM, Magesa S, Mbogo C, Mosha F, Midega J, Lomas S *et al.* (2014). Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. *Mol Biol Evol* **31**: 889–902.

Ometto L, Stephan W, De Lorenzo D (2005). Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* **169**: 1521–1527.

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL (2014). Positive selection drives faster-Z evolution in silkmoths. *Evolution* **68**: 2331–2342.

Sanchez-Gracia A, Rozas J (2007). Unusual pattern of nucleotide sequence variation at the *OS-E* and *OS-F* genomic regions of *Drosophila simulans*. *Genetics* **175**: 1923.

Schlenke TA, Begun DJ (2004). Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc Natl Acad Sci USA* **101**: 1626–1631.

Slater G, Birney E (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D (2012). Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse *(Mus musculus)*. *PLoS Genet* **8**: e1002891.

Tay W, Behere G, Heckel D, Lee S, Batterham P (2008). Exon-primed intron-crossing (EPIC) PCR markers of *Helicoverpa armigera* (Lepidoptera: Noctuidae). *Bull Entomol Res* **98**: 509–518.

Tay WT, Soria MF, Walsh T, Thomazoni D, Silvie P, Behere GT *et al.* (2013). A brave new world for an Old World pest: *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Brazil. *PLoS One* **8**: e80134.

Vicoso B, Charlesworth B (2006). Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet* **7**: 645–653.

Wang C, Dong J (2001). Interspecific hybridization of *Helicoverpa armigera* and *H. assulta* (Lepidoptera: Noctuidae). *Chinese Sci Bull* **46**: 489–491.

Willingham AT, Keil T (2004). A tissue specific cytochrome P450 required for the structure and function of *Drosophila* sensory organs. *Mech Dev* **121**: 1289–1297.

Wondji CS, Hemingway J, Ranson H (2007). Identification and analysis of single nucleotide polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector. *BMC Genomics* **8**: 1–13.

Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z *et al.* (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm *(Bombyx)*. *Science* **326**: 433–436.

Yasukochi Y, Ashakumary LA, Baba K, Yoshido A, Sahara K (2006). A second-generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects. *Genetics* **173**: 1319–1328.

Zalucki M, Daglish G, Firempong S, Twine P (1986). The biology and ecology of *Heliothis armigera* (Hubner) and *Heliothis punctigera* Wallengren (Lepidoptera, Noctuidae) in Australia: what do we know? *Aust J Zool* **34**: 779–814.

Zhang DX (2004). Lepidopteran microsatellite DNA: redundant but promising. *Trends Ecol Evol* **19**: 507–509.

Zhou X, Faktor O, Applebaum SW, Coll M (2000). Population structure of the pestiferous moth *Helicoverpa armigera* in the Eastern Mediterranean using RAPD analysis. *Heredity* **85**: 251–256.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)

2.3 Supplementary material

Locus	Number of pairwise comparisons	Significant pairwise comparisons by Fisher's exact test	After Bonferroni correction	Significant pairwise comparisons by χ^2 test	After Bonferroni correction
Apt	990	286	83	301	202
Cycle	300	86	15	131	18
Cyp303a1	1830	1119	683	1156	633
Cyp305b1	741	97	5	126	43
Period	378	40	1	48	30
Phc	1953	325	28	397	115
SCAP	780	93	0	171	0
Tc	946	110	0	129	0
Tpi	2926	375	36	549	87

Table S1 Average r^2 and number of significant pairwise comparisons before and after Bonferroni correction

Locus	Accession		Primer sequence
Apt	AB024903.1	forward:	CTTGGTCTACTGCCGACCTCACT
		reverse:	TAGCCATCCTCATTGTTTGGACT
Cycle	BAB20632.1	forward:	GCTATGGCGAGGAAGTTAGACAA
		reverse:	TCTCTCCCGTGGGCTGAGGT
Cyp303a1	NM_143813.2	forward:	ACGGCATTTCATGGGGCGCA
		reverse:	GTGCTTGGCTAGGCCGAACGGA
		forward, nested:	GCTCGCTGGATATTTTATACCAGA
Cyp305b1	NP_001106220.1	forward:	ACACGTCTGCGTTTCTCCAA
		reverse:	GGTGTAGCCAATATACCAATCAAC
down3		forward:	TCTCCATTCGGTTTCCCTTC
		reverse:	TAATCACTGGGTTGCTTCTGG
Period	ABF21088.1	forward:	CAGGGACTCGGGTGAGATGA
		reverse:	AGCACTGGTTGGATGGTAGG
Phc	XP_002432072.1	forward:	TACGCGAAGATGTGGTACAAGG
		reverse:	ACAAGTCCATCGGCGGTCTG
SCAP	XP_974766.1	forward:	AAAGCCAAAGCGAGCATAGCA
		reverse:	TACTGAACAAGCAGCCAGACC
Tc	XP_972068.2	forward:	TGTATTCCCAAGTCCGCTGTTT
		reverse:	TTGTTGATAGCTTCGCAAGAGT
Tpi	AY736358	forward:	ATTCGTTGTTGGTGGTAACTGGA
		reverse:	TTTGCCTGCCTCCCTCTCTTCT
up3		forward:	GTTGCGAGTAAGATAGCAGCAC
		reverse:	GTTGCGTGGCAGGAAGGTAA

Table S2 Accessions of Z-linked loci and primer sequences used in this study. In cases where sequence quality was unsatisfactory, an internal (nested) primer was used for re-sequencing. The loci *down3* and *up3* refer to the regions 3 kb downstream and upstream of *Cyp303a1*, respectively, in which no known genes were identified.

	F_{ST} values for pairwise comparisons of populations		
Locus	**Dal X Orb**	**Dal X MV**	**Orb X MV**
Apt	-0.077	0.020	0.013
Cyp303a1	-0.036	-0.012	0.018
Period	-0.017	-0.005	0.025
Phc	0.059	-0.028	-0.009
Tpi	-0.044	0.003	-0.176

Table S3 F_{ST} values for five loci sampled from Dalmore, Orbost and MacIntyre Valley populations using an unbiased estimator (Hudson *et al.*, 1992)

Locus	n	Number of sites (bp)[a]	Number of indel events		π(i)	
Apt **(55)**	Dalmore (8)	779–840	8	23	0.004	0.005
	Orbost (5)		4		0.003	
	M.Valley (41)		20		0.005	
Cycle **(21)**	M.Valley (20)	919		11		0.002
Cyp303a1 **(83)**[b]	Dalmore (14)	774–780	13	20	0.005	0.006
	Orbost (10)		7		0.005	
	M.Valley (56)		17		0.006	
Cyp305b1 **(22)**	M.Valley (21)	696		8		
Period **(36)**	Dalmore (13)	486–502	27	46	0.015	0.014
	Orbost (5)		12		0.011	
	M.Valley (17)		32		0.013	
Phc **(36)**	Dalmore (12)	723–754	28	38	0.011	0.010
	Orbost (7)		20		0.010	
	M.Valley (16)		25		0.009	
SCAP **(12)**	M.Valley (11)	974		20		
Tc **(16)**	M.Valley (15)	914		16		0.004
Tpi **(33)**	Dalmore (11)	551–577	18	21	0.009	0.007
	Orbost (2)		3		0.005	
	M.Valley (19)		14		0.005	

Table S4 Indel diversity for nine loci surveyed in this study. Where estimates are presented in two columns under a single heading, the left column represents estimates for an individual population while the right column represents estimates after pooling sequences of all three populations. Figures in brackets after the locus name represent the total number of sequences surveyed including the reference strain.

* *$p<0.05$*

[a] *the number of sites is presented as a range due to the differing subsets of indel polymorphisms present in different populations. Since gapped sites are included in this analysis, the upper boundary represents the number of sites considered when alleles from all three populations are pooled.*

[b] *includes two sequences from a laboratory-maintained colony*

Site	*Del200*	*Ins200*	*H. assulta*	*H. punctigera*	*H. zea*
1	ATAA	-	x	x	GTAA
2	G(1A)	C	C	x	C
3	C	T	T	x	T
4	T	C	C	x	T
5	G	A/T	T	x	T
6	T(1G)	A	A/C	x	A
7	C	G	G	x	G
8	A	G	G	x	G
9	G	A	G	x	G
10	C(1G)	A	C	x	C
11	G	T	T	x	T
12	G	C	G/A	x	C
13	T	A	T	x	T
14	G	T	G	x	T
15	A	G	G	x	G
16	T	C	T/-	x	C
17	G	T	T/-	x	T
18	C	G	C	x	C
19	C	A	C	x	C
20	A	G	G	x	-
21	A	G	G	x	-
22	A	G	A	x	G
23	A	C	C	x	C
24	T	A/C	T	T	T
25	A	C	C	C	C
26	A	T	T	T	T
27	T	C	C	C	C
28	C	A	A	A	C
29	T	A	A	A	G
30	AACG	-	AATG	AATN	-
31	-	A	GA	-	A
32	A	C	A	A	G

33	A	C	A	A	G
34	A	G	A	A	T
35	G	A	A	A	A
36	G	A	AG	-	G
Number of sites with state identical or closer to the					
Del200 haplogroup			10	14	5
Ins200 haplogroup			20	19	6
Ambiguous unknown:			6	3	25

Table S5 State of the outgroup consensus sequences at sites where fixed differences occur between the *H. armigera* insertion and deletion haplogroups. '-' represents a gap in the alignment flanked by matching nucleotides whereas 'x' represents a gap that forms part of a deletion extending beyond the boundaries of the sites where the fixed differences occur. All sites marked with an 'x' are deemed ambiguous. Sites with a '-' are considered unambiguous if they match one of the two states in the *H. armigera* alleles e.g. site 20 in *H. zea* is considered ambiguous but site 30 is not. Sites which require more than one change from either allele are also deemed ambiguous e.g. site 30 in *H. assulta* is considered to be closer to the insertion state but the same site in *H. punctigera* is deemed ambiguous because a single event is insufficient to produce a match to either state. Similarly, site 31 in *H. assulta* is ambiguous because only one event (a mono- or di-nucleotide insertion) is sufficient to produce a match to either state. Where polymorphisms are present in the outgroup sequences, they are deemed unambiguous if at least one of the states is identical to that of *H. armigera* and all possible states are observed to be exclusive of the alternate *H. armigera* allele. Sites 2, 5, 6, 10 and 24 are, strictly speaking, not fixed differences, but have been included because no overlapping states have been observed between the two *H. armigera* alleles.

Organism	π (10^2)	Approximate distance (bp) at which E(r^2)≈0.2	References
Lepidoptera			
Bombyx mori	1	>1600[a]	Xia *et al.* (2009); Guo *et al.* (2011)
Bombyx mandarina	1–2	20–200[a]	Xia *et al.* (2009); Guo *et al.* (2011)
Heliconius erato	2	<500	Counterman *et al.* (2010)
Heliconius melpomene	1	>500	Baxter *et al.* (2010)
Helicoverpa armigera	3	10–200[b]	this study
Melitaea cinxia	n.a	3000	Ahola *et al.* (2014)
Other insects			
Acyrthosiphon pisum	0.6	1000	Brisson *et al.* (2009)
Anopheles arabiensis	0.2–0.3	<200	Marsden *et al.* (2014)
Anopheles gambiae	0.8–25	<200	Wilding *et al.* (2009); Harris *et al.* (2010)
Drosophila melanogaster	0.6–2	30[c]–640[d]	Langley *et al.* (2012); Pool *et al.* (2012)

Table S6 Estimates of nucleotide diversity, π and linkage disequilibrium, E(r^2) in different species. Unless otherwise indicated, LD was imputed from diploid sequences.

[a] *haploid by cloning amplicons prior to sequencing*

[b] *haploid by sequencing sex chromosomes in hemizygous individuals*

[c] *haploid embryos*

[d] *haploid by using inbred lines*

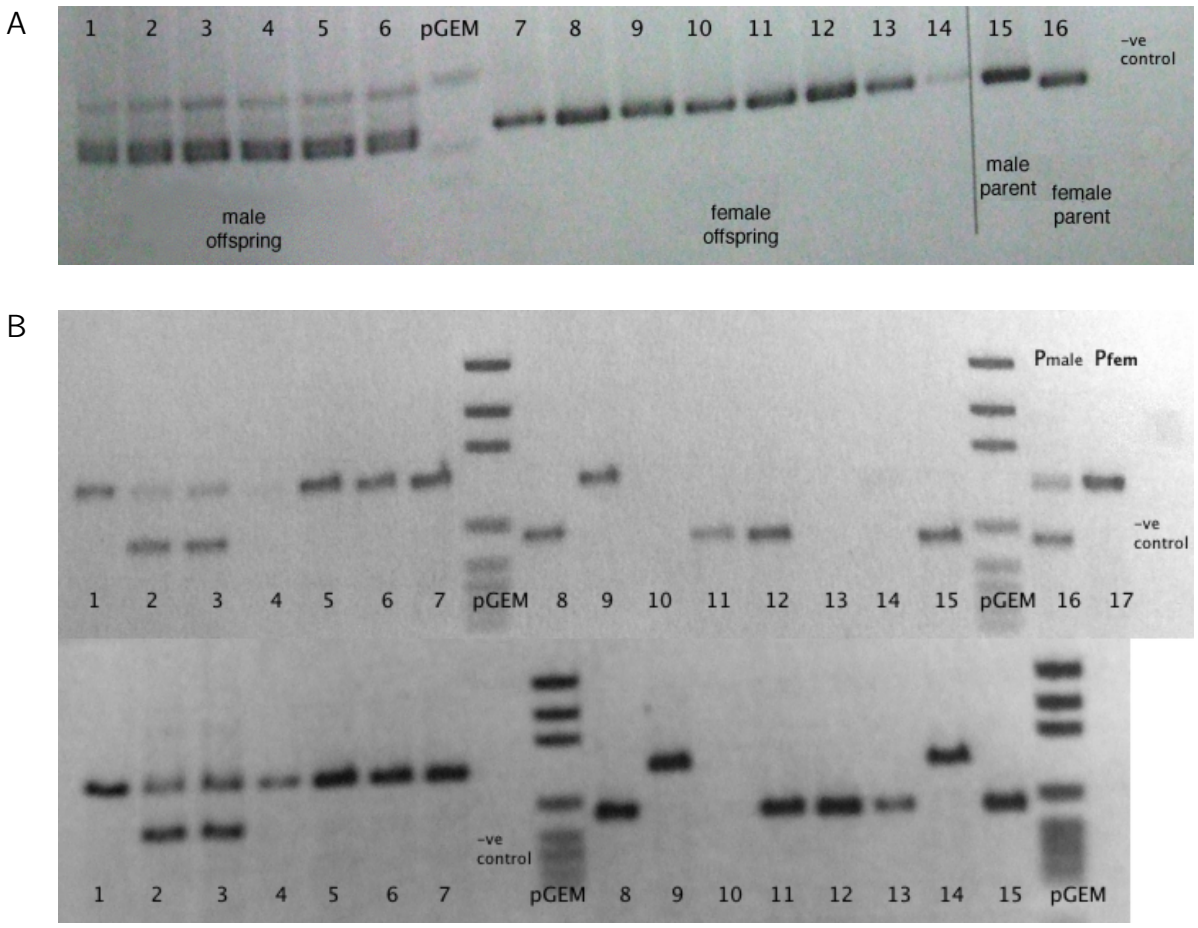


Figure S1 Pedigree analysis of (A) *Period*. All male offspring (lanes 1–6) present double bands inherited from both parents while all female offspring (7–14) inherit a single band from the male parent (15). The female parent has a slightly shorter product (16) and her allele is only seen in male offspring. Pedigree analysis of (B) *Phc*. The upper gel shows male offspring (lanes 1–7), female offspring (8–15), male parent (16) and female parent (17). The male parent is heterozygous. The lower gel stems from a technical replicate (different PCR) to account for lack of product in samples 4, 10, 13 and 14. Male offspring exhibit heterozygosity (2,3) while female offspring (excluding sample 10) exhibit hemizygosity.

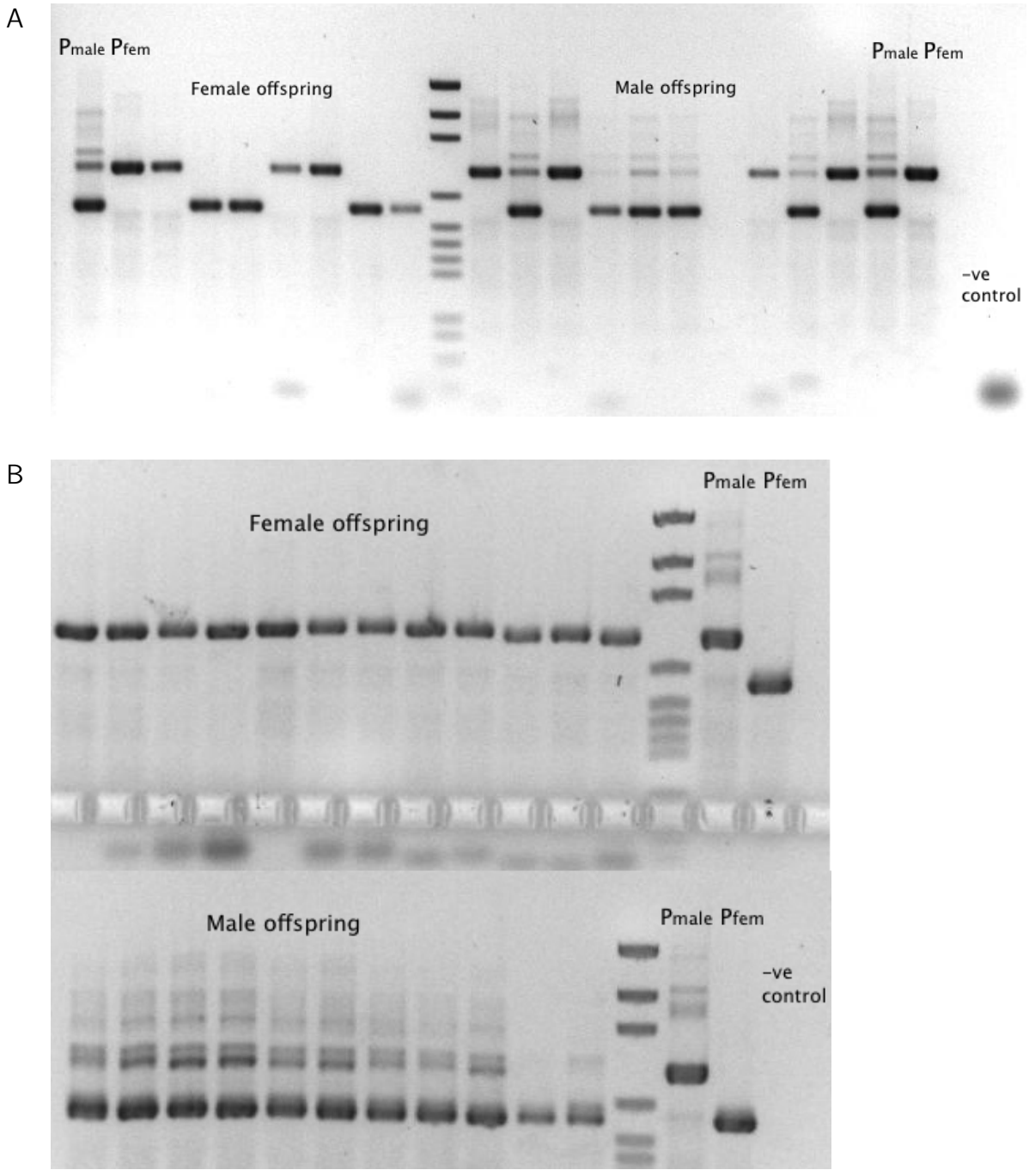


Figure S2 Pedigree analysis on the *Cyp303a1* insertion and deletion variants. The male parent of family 1 (A) is heterozygous while that of family 2 (B) is homozygous for the insertion. In both cases, female offspring exhibit hemizygosity whereas male offspring may be homozygous or heterozygous.

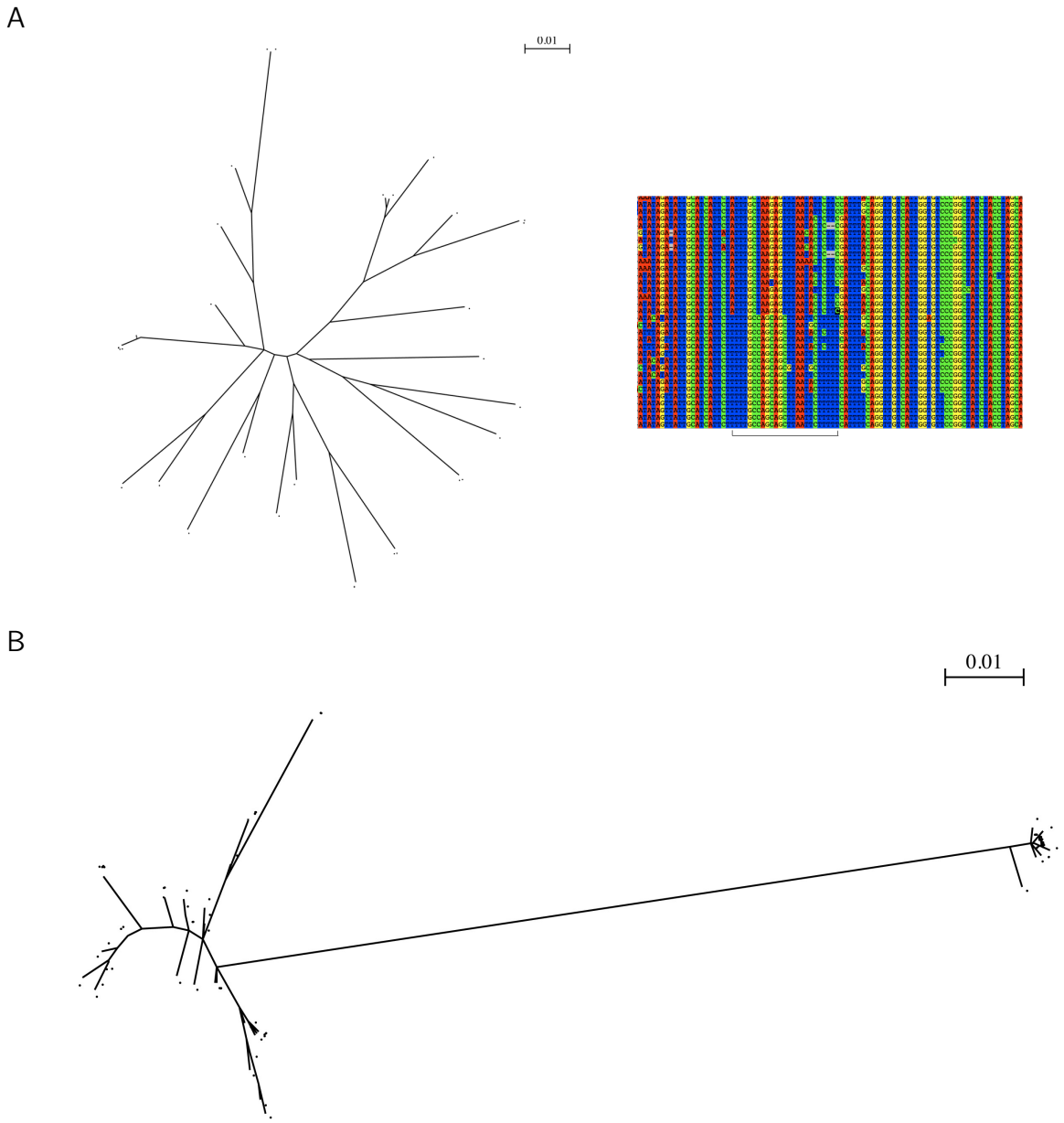
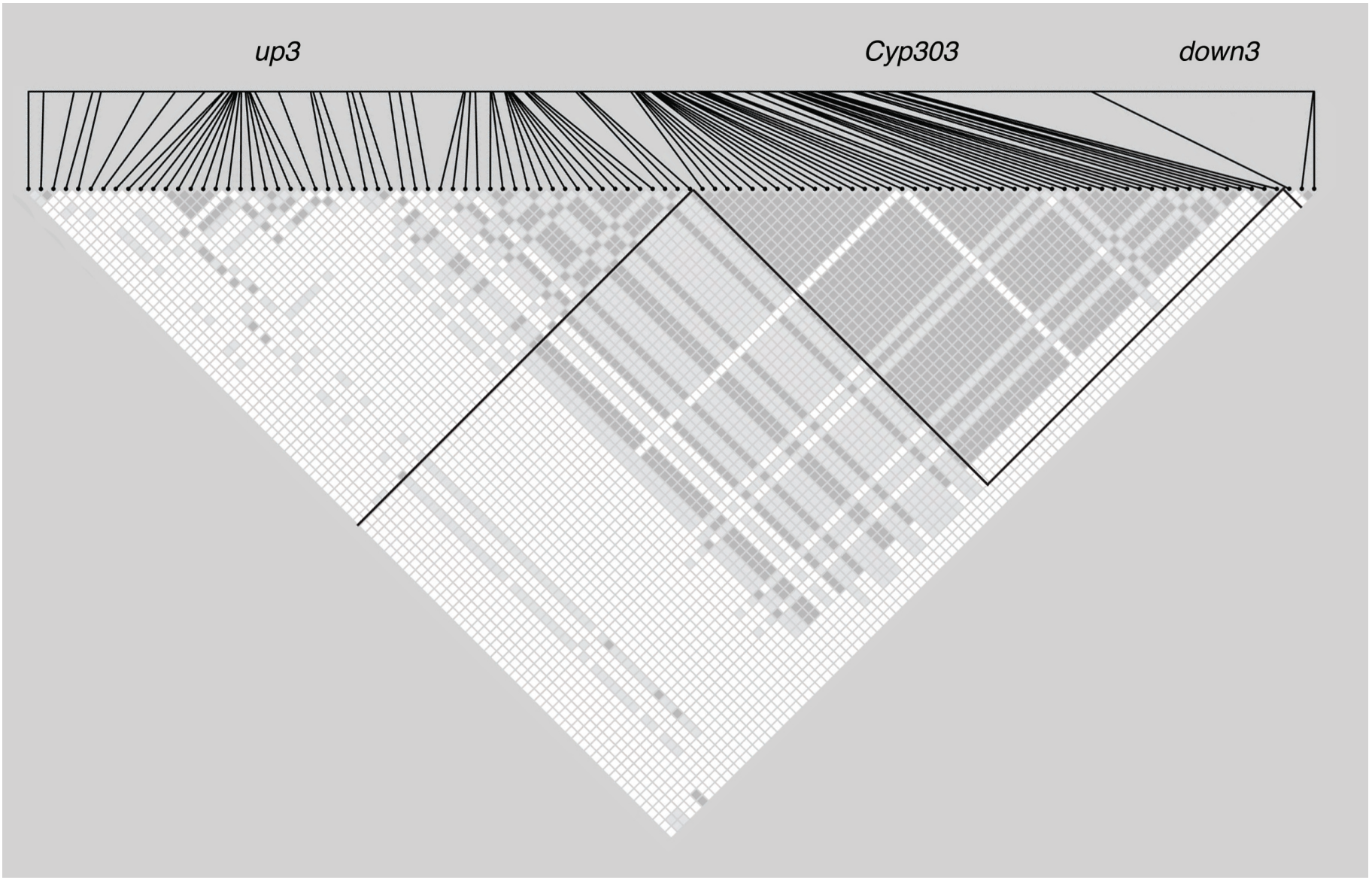


Figure S3 Unrooted maximum-likelihood trees for (A) *Tpi* and (B) *Cyp303a1* after exclusion of the 25-bp stretch and 200-bp indel, respectively. The long internal branch is still apparent in *Cyp303a1*. The fixed differences between the insertion and deletion alleles cannot be accounted for by a consecutive stretch of nucleotides that would constitute a single event. In contrast, the alignment in (A) illustrates that excluding a single 25-bp stretch in *Tpi* (underlined with black bar) largely eliminates the long internal branch.


up3
Cyp303
down3

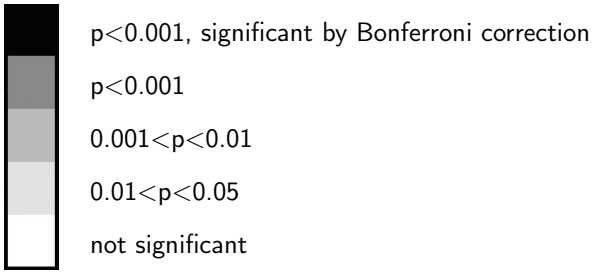


Figure S4 ***(preceding page)*** LD plotted as significance of the r^2 value for *Cyp303a1* and its flanking regions. The sequence under consideration was generated by concatenating the *up3*, *Cyp303a1* and *down3* loci (n=11). The range of each locus is bounded by black lines. Approximately 3 kb separate *Cyp303a1* from its flanking sequences. The downstream locus contained very few segregating sites, hence the small area. LD is observable between the 3' end of the upstream locus and *Cyp303a1*.

SUPPLEMENTARY REFERENCES

Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Valimaki N, Paulin L, Kvist J, Wahlberg N, Tanskanen J, Hornett EA, Ferguson LC, Luo S, Cao Z, de Jong MA, Duplouy A, Smolander OP, Vogel H, McCoy RC, Qian K, Chong WS, Zhang Q, Ahmad F, Haukka JK, Joshi A, Salojarvi J, Wheat CW, Grosse-Wilde E, Hughes D, Katainen R, Pitkanen E, Ylinen J, Waterhouse RM, Turunen M, Vaharautio A, Ojanen SP, Schulman AH, Taipale M, Lawson D, Ukkonen E, Makinen V, Goldsmith MR, Holm L, Auvinen P, Frilander MJ and Hanski I (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun* **5**

Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A, Beltran M, Perez-Espona S, Chamberlain N, Ferguson L, Clark R, Davidson C, Glithero R, Mallet J, McMillan WO, Kronforst M, Joron M, ffrench Constant RH and Jiggins CD (2010). Genomic hotspots for adaptation: The population genetics of Mullerian mimicry in the *Heliconius* melpomene clade. *PLoS Genet* **6**: e1000794

Brisson J, Nuzhdin S and Stern D (2009). Similar patterns of linkage disequilibrium and nucleotide diversity in native and introduced populations of the pea aphid, *Acyrthosiphon* pisum. *BMC Genetics* **10**: 22

Counterman BA, Araujo-Perez F, Hines HM, Baxter SW, Morrison CM, Lindstrom DP, Papa R, Ferguson L, Joron M, ffrench Constant RH, Smith CP, Nielsen DM, Chen R, Jiggins CD, Reed RD, Halder G, Mallet J and McMillan WO (2010). Genomic hotspots for adaptation: The population genetics of Mullerian mimicry in *Heliconius* erato **6**

Guo Y, Shen YH, Sun W, Kishino H, Xiang ZH and Zhang Z (2011). Nucleotide diversity and selection signature in the domesticated silkworm, *Bombyx* mori, and wild silkworm, *Bombyx* mandarina. *Journal of Insect Science* **11**: 155

Harris C, Rousset F, Morlais I, Fontenille D and Cohuet A (2010). Low linkage disequilibrium in wild *Anopheles* gambiae s.l. populations. *BMC Genetics* **11**: 81

Hudson RR, Slatkin M and Maddison WP (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, Fang S, Nista PM, Holloway AK, Kern AD, Dewey CN, Song YS, Hahn MW and Begun DJ (2012). Genomic variation in natural populations of *Drosophila* melanogaster. *Genetics* **192**: 533–598

Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, Ferguson HM, Eskin E and Lanzaro GC (2014). Diversity, differentiation, and linkage disequilibrium: prospects for association mapping in the malaria vector Anopheles arabiensis. *G3 (Bethesda)* **4**: 121–131

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ and Langley CH (2012). Population genomics of sub-Saharan *Drosophila* melanogaster: African diversity and non-African admixture. *PLoS Genet* **8**: e1003080

Wilding C, Weetman D, Steen K and Donnelly M (2009). High, clustered, nucleotide diversity in the genome of *Anopheles* gambiae revealed through pooled-template sequencing: implications for high-throughput genotyping protocols. *BMC Genomics* **10**: 320

Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z, Dai F, Li Y, Cheng D, Li R, Cheng T, Jiang T, Becquet C, Xu X, Liu C, Zha X, Fan W, Lin Y, Shen Y, Jiang L, Jensen J, Hellmann I, Tang S, Zhao P, Xu H, Yu C, Zhang G, Li J, Cao J, Liu S, He N, Zhou Y, Liu H, Zhao J, Ye C, Du Z, Pan G, Zhao A, Shao H, Zeng W, Wu P, Li C, Pan M, Li J, Yin X, Li D, Wang J, Zheng H, Wang W, Zhang X, Li S, Yang H, Lu C, Nielsen R, Zhou Z, Wang J, Xiang Z and Wang J (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm *(Bombyx*). *Science* **326**: 433–436

Chapter 3

Population differentiation between Australian and Chinese *Helicoverpa armigera* occurs in distinct blocks on the Z chromosome

3.1 Introduction

The primary goal of this paper is to assess the extent of population structure between *H. armigera* from different continents, and to identify candidate loci that are informative in that respect. This study extends the aims of the previous study by assessing whether or not the results (estimates of nucleotide diversity and linkage disequilibrium) obtained in Paper 1 can be generalised to other populations of *H. armigera*. These parameters are reported and subsequently used to quantify the differences between populations. We then use a chromosome-wide scan to identify regions of high differentiation between Australian and Chinese *H. armigera* and characterise loci that may be informative for inter-continental population structure. As *H. armigera* is a globally-distributed pest of economically important crops, understanding the extent of gene flow has implications for managing the spread of resistance alleles in the field. This paper thus provides applied pest researchers with some useful markers and recommendations for characterising population structure in this species.

A secondary aim is to characterise how patterns of differentiation are distributed across the Z chromosome, and to assess the evidence for competing hypotheses of how such differentiated regions arose in the genome. We also assess the evidence for the hypothesis that Australian *H. armigera* may belong to a subspecies distinct from that of non-Australian *H. armigera*. Questions about species boundaries remain an active area of inquiry in biology, and this paper explores the idea that highly-differentiated regions in the *H. armigera* genome may have been a result of introgressive events.

3.2 Paper 2

TITLE:

Population differentiation between Australian and Chinese *Helicoverpa armigera* occurs in distinct blocks on the Z-chromosome

RUNNING TITLE:

Z chromosome variation in *Helicoverpa armigera*

KEYWORDS:

CYP303A1, *Helicoverpa armigera conferta*

AUTHORS:

Sue Vern Song[1,2], Craig Anderson[3,4], Robert Trygve Good[1,2], Stephen Leslie[1,5,6], Yidong Wu[7], John Graham Oakeshott[4], Charles Robin[1,2]

1. School of Biosciences, University of Melbourne, Victoria, Australia.

2. Bio21 Institute, Parkville, Victoria, Australia.

3. Biological and Environmental Sciences, University of Stirling, Stirling, FK9 4LA, United Kingdom.

4. Land and Water Flagship, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australian Capital Territory, Australia.

5. School of Mathematics and Statistics, University of Melbourne, Victoria, Australia.

6. Centre for Systems Genomics, University of Melbourne, Victoria, Australia.

7. College of Plant Protection, Nanjing Agricultural University, Nanjing, China

ABSTRACT

Over the last 40 years, many types of population genetic markers have been used to assess the population structure of the pest moth species *Helicoverpa armigera*. While this species is highly vagile, there is evidence of inter-continental population structure. Here, we examine Z chromosome molecular markers within and between Chinese and Australian populations. Using 1352 polymorphic sites from 40 Z-linked loci, we compared two Chinese populations of moths separated by 700km and found virtually no population structure (n=41 and n=54, with less than 1% of variation discriminating between populations). The levels of nucleotide diversity within these populations were consistent with previous estimates from introns in Z-linked genes of Australian samples (π=0.028 versus 0.03). Furthermore, all loci surveyed in these Chinese populations showed a skew towards rare variants, with ten loci having a significant Tajima's D statistic, suggesting that this species could have undergone a population expansion. Eight of the 40 loci had been examined in a previous study of Australian moths, of which six revealed very little inter-continental population structure. However, the two markers associated with the *Cyp303a1* locus that has previously been proposed to be a target of a selective sweep, exhibited allele structuring between countries. Using a separate dataset of 19 Australian and four Chinese moths, we scanned the molecular variation distributed across the entire Z chromosome and found distinct blocks of differentiation that include the region containing *Cyp303a1*. We recommend some of these loci join those associated with insecticide resistance to form a set of genes best suited to analyzing population structure in this global pest.

INTRODUCTION

Helicoverpa armigera is one of several polyphagous moths classified within the so-called 'mega pest' lineage of heliothine moths which also includes *Chloridea virescens*, *H. zea* and *H. punctigera*. It causes billions of dollars of crop damage annually. The genus Helicoverpa most probably evolved within Australasia, as the most basal lineage within it (*Helicoverpa punctigera*) is only found in Australia, and the closest sister genus, *Australothis* is also found there (Matthews, 1999; Cho *et al.*, 2008). Five (*H. armigera*, *H. assulta*, *H. hardwicki*, *H. prepodes* and *H. punctigera*) of the approximately 20 *Helicoverpa* species described are found in Australia (Gordon *et al.*, 2010; Mitchell & Gopurenko, 2016). It is less clear where the species *H. armigera* arose. A typical way of inferring origins, the centre of nucleotide diversity, does not provide a clear answer in the literature published thus far (Nibouche *et al.*, 1998; Zhou *et al.*, 2000; Behere *et al.*, 2007). The age of the *H. armigera* diaspora is also somewhat clouded, although mounting evidence suggests it arose after the divergence from the New World sibling species, *H. zea*, which has been estimated as 1.5 – 2 mya (Mallet *et al.*, 1993; Behere *et al.*, 2007; Pearce *et al.*, 2017).

Until recently, the distribution of *H. armigera* was limited to Europe, Asia, Africa and Australia. However, the species was detected in South America around 2012 and is currently spreading through the Americas, causing significant damage (Czepak *et al.*, 2013; Tay *et al.*, 2013; Murua *et al.*, 2014; Arnemann *et al.*, 2016; Sosa-Gomez *et al.*, 2016). Many countries currently control this pest with strategies that combine crops genetically modified to produce *Bacillus thuringiensis* (*Bt*) insecticidal proteins with practices such as the planting of 'refuges' aimed to slow the spread of recessive resistance alleles (Tabashnik *et al.*, 2004). The recent range expansion of *H. armigera* into areas likely to have heterogeneous insecticide exposures and control strategies may increase the likelihood of insecticide resistance alleles

proliferating. Resistant alleles may then flow back to populations in the ancestral range. Thus tracking gene flow of insecticide resistance alleles is important to control this damaging species throughout its range (Daly, 1993; Fitt, 1994).

Population genomics offers a new way to identify insecticide resistance alleles because they may be associated with selective sweeps. Under the scenario where a resistance allele reaches high frequency (or fixation) in a population, the polymorphisms in neighbouring gene regions will exhibit atypical patterns such as reduced within-species variation, elevated levels of differentiation among populations, extended haplotypes and a skewed frequency spectrum such that a greater fraction of the variants are rare (Nielsen, 2005). Studies in *Drosophila melanogaster* underscore the potential of this approach, with genome-wide scans for positive selection successfully recovering strong signals at known resistance loci (Garud *et al.*, 2015; Battlay *et al.*, 2016) despite the fact that *D. melanogaster* is not itself a direct target of insecticides. Thus these findings hold much promise for organisms such as *H. armigera*, which would be expected to face much stronger selective pressures as targeted pest species.

As high-throughput sequencing technologies continue to be refined, increasingly sophisticated approaches are available to reduce the complexities in genomic data yet capture the inherent architectures that are unique to each species. These reduced-representation sequencing technologies include RAD-Seq (Baird *et al.*, 2008; Rasic *et al.*, 2014) and genotyping-by-sequencing (GBS; Elshire *et al.*, 2011). Recently, Anderson *et al.* (2016) investigated worldwide population genomic variation in *H. armigera* with two different ways of sampling alleles from portions of the genome. In one, they examined 21,043 SNPs located across the genome among 216 individuals using a GBS approach. In the other, they examined the variation among 50 individuals in specific fully sequenced regions aligning to 2.3Mb of 20 BAC clones. They found clear evidence for inter-continental population structure in their full

mitochondrial sequence dataset and in their GBS and BAC-aligned sequence datasets. This prompted them to resurrect an idea that *H. armigera* from Australia had its own subspecies, named *H. armigera conferta*, that differs from *H. armigera armigera* in Africa, Europe and Asia (Common, 1953; Matthews, 1999). Several groups have used mtDNA to examine population structure, but these earlier datasets show a shallow star phylogeny and little structure (Behere *et al.*, 2007; Tay *et al.*, 2013; Leite *et al.*, 2014; Mastrangelo *et al.*, 2014). Microsatellite studies have also been used to examine this issue but interpretations are confounded by non-Mendelian patterns of the markers attributable to a high frequency of null alleles and to associations with transposable elements (Zhang *et al.*, 2004; Tay *et al.*, 2010). Behere *et al.* (2013) primarily used autosomal exon-primed intron-crossing (EPIC) markers to address population structure issues between crops in India, and they too showed little compelling evidence for population structure.

Our previous study of the Z chromosome of Australian *H. armigera* indicated that the amount of variation observed within populations was very high relative to other organisms (nucleotide diversity, π~0.02) and that recombination between sites must also be high as linkage disequilibrium (LD) decayed rapidly, with E(r^2) approaching 0.2 within 200bps (Song *et al.*, 2015). Given that these metrics will differ due to the varying degrees of influence that neutral and non-neutral processes have on a finite population, we wanted to explore the robustness of the estimated parameters and examine if these observations could be extended to non-Australian populations. *H. armigera* follows a ZZ/ZW sex determination system with females being the heterogametic sex. As in Song *et al.* (2015), a crucial aspect of the design we follow here is the focus on Z-linked loci because sequencing females provides empirical observations of haplotypes and overcomes the limitations of imputation solutions to the gametic-phase problem (in an individual heterozygous at two loci, the gametic phase is either AB/ab or Ab/aB and which one it is affects LD calculations) which are particularly inaccurate when LD

decays rapidly (Slatkin, 2008).

Here, we use a Z-linked EPIC dataset and a Z chromosome-wide scan to identify candidate loci that are informative for inter-continental population structure and to assess the *conferta* sub-species hypothesis. The first dataset employs targeted resequencing to examine nucleotide diversity and LD at 40 EPIC markers in female samples from two collection sites in China. The use of 454 pyrosequencing of PCR amplicons provides us with relatively long reads (~600bp), and by analyzing only female moths which contain only a single Z chromosome, we cleanly sample a single allele per individual. The 40 markers include eight loci characterized in Australian samples from our previous study (Song *et al.*, 2015), which allows us to explore the question of differentiation between inter-continental populations of *H. armigera.* The second dataset uses previously reported whole-genome sequencing from Anderson *et al.* (2016) to scan the Z chromosome for regions that show high levels of differentiation between Australian and Chinese individuals. This dataset provides us with an independent sample of moths from the two countries to investigate how patterns of diffentiation are distributed across the chromosome whilst further exploring the question of inter-continental population structure.

MATERIALS AND METHODS

DNA samples

Our analysis of polymorphic sites in the 40 Z-linked EPIC (exon-primed intron-crossing) markers in the Chinese samples was based on 41 female adults from Nanpi (Hebei Province, Yellow River Valley) (38°2′N, 116°42′E) and 54 female adults from Yancheng (Jiangsu Province, Changjiang River Valley) (33°20′N, 120°9′E) collected in 2011. DNA was isolated by column purification. The provenance of the Australian samples has been reported in Song *et al.* (2015). Briefly, they consist of eggs collected from MacIntyre Valley (Queensland) which were reared to adults in the laboratory, after which only female samples were used.

Samples used in the Z-chromosome-wide analysis comprised 19 Australian and four Chinese individuals used in Anderson *et al.* (2016). In this case, *H. armigera* were collected as adults using emergence traps placed among cotton plants in New South Wales, Australia, and as caterpillars from cotton in Shandong, China. DNA was isolated by column purification.

Library preparation

For the EPIC markers, Z-linked loci were identified in the silkwork *Bombyx mori* using the published GLEAN cDNA dataset for this species and coding sequences were extracted based on the Z chromosome (chromosome 1) scaffold numbers (nscaf1690, nscaf2210, nscaf2734, nscaf3040 and nscaf3068 from http://silkworm.genomics.org.cn/). The *H. armigera* orthologues of these *B. mori* loci were identified though BLAST searches against a repository containing contigs from an *H. armigera* reference strain (Pearce *et al.,* 2017). Intron-exon boundaries on the *H. armigera* contig sequences were identified with EXONERATE under the *cdna2genome* model using the silkworm cDNA as input queries. Loci were selected to

include different regions of the Z chromosome and to incorporate clusters of loci, where the distance separating two loci was less than 50kb (Figure 3.1) to allow for the possibility of detecting long-range LD.

The 454 Universal Tailed Amplicon (Roche) sequencing design was used where sequences from each individual were given unique barcodes. Briefly, the strategy utilizes two successive rounds of PCR to produce amplicons appropriate for the sequencing platform. The first round of PCR employs fusion primers consisting of the universal tail sequence and a gene-specific primer. The second round of PCR is carried out with fusion primers comprising the 454 sequencing primers, a custom 5-bp MID (multiplex identifier) sequence, and a sequence which targets the tail sequences introduced in the first round. Primer sequences are provided in Supplementary Table S1 and loci have been named according to the *B. mori* gene they are based on; those with the suffix BGIBMGA refer to *Bombyx mori* orthologs using the names assigned by the silkworm sequencing consortium (http://silkworm.genomics.org.cn).

First-round amplification with locus-specific primers was carried out in a reaction volume of 20µl. Cycling conditions were typically 35 cycles of 94°C for 20 seconds, 58°C for 20 seconds, 68°C for 1 minute. To estimate the size and intensity of bands, 3µl of each product was visualized on agarose gels with 5µl of DNA markers. Each amplicon was assigned a score of intensity ranging from 2 (strongest) to 10 (weakest) prior to pooling by sample barcodes (MIDs) for library preparation and sequencing. To compensate for different yields (to avoid over-sampling of particular amplicons during sequencing), pooling was carried out in the following manner: 2µl of an amplicon was included if it had a score of 2, and up to 10µl was included for weaker products.

Bead purification was carried out with MagNA buffer (Rohland & Reich, 2012) and 1µl of each pool was used as a template for second-round amplification in individual reaction

volumes of 50µl. Second-round amplification was carried out in duplicate with 10 cycles of 94°C for 30 seconds, 68°C for 1.5 minutes to incorporate the MID barcodes and adapter sequences. The barcoded pools were combined into a single library then purified and concentrated by bead purification, gel excision and column purification prior to quantification. Sequencing was performed by The Ramaciotti Centre for Genomics (University of New South Wales, Australia) using Roche 454 GS FLX Titanium chemistry for amplicon sequencing (XLR70).

For the Z-chromosome-wide dataset, library preparation has been described in Anderson *et al.* (2016). Briefly, Nextera libraries were produced following the manufacturer's instructions, and sequence was generated as 100-bp PE reads (Illumina HiSeq 2000, Biological Resources Facility, Australian National University, Australia).

Sequence diversity and linkage disequilibrium

For the EPIC dataset, sequence reads were first sorted by MIDs (individual barcodes) and assembled *de novo* in GENEIOUS R7 (Biomatters, Auckland, New Zealand) using a threshold of 10% for maximum mismatches per read. Contigs from all individuals were then mapped to the reference *H. armigera* amplicon sequences using BLASTN. Mapped contigs comprised 116,069 reads, representing 83% of the total number of reads. Multiple sequence alignments for each locus were performed using SEAVIEW 4.0 (Gouy *et al.*, 2010) and CLUSTALX (Larkin *et al.*, 2007). Sequences were assessed at two levels in the process of error correction: at the individual (MID) level and at the population (aligned to reference) level. Where low coverage at the individual level resulted in base ambiguities, this was resolved by comparing to other sequences in the population i.e. if all other individuals were

monomorphic at that site, the ambiguity was manually edited to match the population. Base ambiguities observed at segregating sites were edited to match the majority (>50%) of individuals in the population. This approach was adopted to counteract the possible inflation of rare SNPs in the population which could result in deviations from the expected site frequency spectrum under neutrality. The filtered dataset contained a total of 40 amplicons distributed across 29 *H. armigera* contigs that map to five *B. mori* Z linked scaffolds (Figure 3.1) with an average of 58 individuals per locus.

Analyses of polymorphism and LD were carried out using DNASP 5.10.01 (Rozas *et al.*, 2003) with alignment files indicated as haploid Z-chromosome. Nucleotide diversity (π) and Tajima's *D* were estimated using a total of 2351 sites. A second estimate of π and θ was also made using only the first 100 bases from the 5' ends of the sequence alignments (with indels excluded) to explore the possibility that diversity could be inflated by sequencing errors towards the 3' ends of longer reads (Gilles *et al.*, 2011). LD was estimated as the square of the correlation coefficient, r^2, using only parsimony-informative sites, with decay of LD over physical distance modelled on the expectations of Hill & Weir (1988) and implemented using the nonlinear least-squares (*nls*) function in R (R Core Team, 2014).

The data analyzed in the frequency spectrum tests above included eight haplotypes (spanning four loci) that were highly diverged from other alleles at those loci. Notably, three of the haplotypes came from a single individual collected from Yancheng, MID-95. At two loci where we had sequences from *H. assulta*, the divergent haplotype of MID-95 resembled the *H. assulta* sequences. We thus excluded MID-95 from all analyses on the likely basis that it was an *assulta* specimen.

For the Z-chromosome-wide dataset, data handling has been described in Anderson *et al.* (2016). Briefly, raw sequence reads were aligned to the Z chromosome of the *H. armigera*

genome using BBMAP v.33.43 (http://sourceforge.net/projects/bbmap/). Quality trimming of reads was carried out when at least two consecutive bases fell below Q10, and only uniquely aligning reads were included in the analysis. UnifiedGenotyper in GATK v. 3.3-0 (McKenna *et al.,* 2010) was used to estimate genotypes, implementing a heterozygosity value of 0.01.

Analyses of population structure

AMOVA, F_{ST} and STRUCTURE analyses

Analysis of molecular variance (AMOVA) and F_{ST} calculations were performed on the EPIC dataset using functions implemented in the R packages *adegenet* v2.0.0, *hierfstat* v0.04-22 and *Poppr* v2.2.0 (Goudet, 2005, Jombart & Ahmed, 2011; Kamvar *et al.*, 2014). We used STRUCTURE version 2.3.4 (Pritchard *et al.*, 2000) to assess the degree of population stratification between Nanpi and Yancheng, and between MacIntyre Valley (Australia) and China. Two approaches were initially considered in the definition of a locus: haplotypes, for which the dataset is well suited due to the phased (hemizygous) nature of the sequences; and sites, which increase the number of loci available for analysis. A haplotype is made up of many polymorphic sites within the same amplicon. Using this 'haplotype approach', our dataset thus contains only 40 loci for analysis. The 'site approach' treats each polymorphic site as a locus, and there were 2351 such sites in our dataset. Singleton sites (where only a single instance of a variant allele was observed) were then excluded as they are uninformative with respect to distinguishing between populations, reducing the number of loci to 1352 in the final analysis. In cases where sequences were not available for all individuals, a value of -9 was assigned to denote missing data.

The haplotype approach avoids any possible issues with non-independence between sites due

to LD, but has the disadvantage of collapsing multiple polymorphic sites into a single haplotype, thus reducing the number of loci for analysis and consequently the power to assign individuals to a specific cluster. A second limitation was the high levels of diversity in all populations, leading to an increase in the uncertainty surrounding the clustering of singleton and low-frequency haplotypes. We attempted to reduce the number of singleton haplotypes by grouping together individuals that differed from a 'core haplotype' at polymorphic sites not represented elsewhere (singleton sites) but even so, overall haplotype diversity remained high, and this motivated the site approach. While a STRUCTURE analysis would not typically consider sites within the same amplicon to be independent markers, the rapid decay of LD in *H. armigera* suggested that a valid analysis could be performed without violating the assumptions of independence between loci (Falush *et al.*, 2003). To assess the robustness of the site approach, a series of 'thinned' datasets were generated by randomly selecting 50% of the 1352 sites available, and subjected to the same STRUCTURE analyses as the full dataset (described below). Ten such datasets were generated, all of which showed a similar outcome to the full dataset, indicating that linked SNPs are not over-represented despite the short distances separating each locus (site). Only the results of the 'site' analysis are presented here as they proved to be more informative than the haplotype analysis.

All STRUCTURE analyses were carried out under the model incorporating admixture and independent allele frequencies between populations, without using prior population information. A series of analyses was run using values of K (the number of genetically defined populations, which may be unknown) from 1–5 using the same parameters. Ten replicates were run for each K value with 10,000 iterations for the burn-in period followed by an additional 10,000 iterations after the burn-in. In cases where a choice lay between K=2 and K>2 to explain most of the structure in the data, the Evanno method (Evanno *et al.*, 2005) was used to formally evaluate the most likely K value via STRUCTURE HARVESTER (Earl &

vonHoldt, 2011) by choosing the value of *K* that corresponds to the largest value of ΔK. For the Nanpi-Yancheng comparison where the choice lay between *K*=1 and *K*=2, no formal evaluation was applied as it is not possible to obtain a value of ΔK between *K*=0 and *K*=1; the most likely value of *K* was inferred from the graphical results.

The output of a STRUCTURE analysis is typically presented in the form of a bar chart illustrating the number of distinct populations as well as assignments of individuals to populations (such as those depicted in Figure 3.2). However, these visualizations represent the output of only a single run. We performed an analysis to explore the correlation between a single STRUCTURE run versus the results of 100 runs (Supplementary Figure S1). The results show that there is a moderate to high level of reproducibility in the assignment of individuals to a population across multiple runs. We therefore conclude that in most cases, the graphical output of a single, randomly-selected run is a reasonable representation of the results of multiple runs.

Z chromosome-wide sliding window analysis of weighted F_{ST}

Imputation of missing bases in the Z-chromosome-wide dataset was performed using default parameters in *Beagle* (Browning & Browning, 2007). Linkage disequilibrium (LD)-based pruning was conducted using *Plink* v.1.07 (Purcell *et al.*, 2007) with the command "--indep 50 5 2". Non-negative weighted F_{ST} and Tajima's *D* were calculated across sliding windows using *vcftools* v.0.1.14 (Danecek *et al.*, 2011) where each window contained 500 SNPs. Sliding window analyses were plotted using R and *ggplot2* (Wickham, 2009). A series of 1000 permutations was carried out on the empirical dataset by randomly partitioning the samples into two groups, each containing 19 and four individuals. The same weighted F_{ST} analysis was then carried out on these permuted datasets whereby the 95th percentile of the F_{ST} values obtained in each window was used to plot the grey dots in Figure 3.4. The aim of

this exercise was to assess the extent to which the empirical observations deviated from the expected distribution of F_{ST} values, given the small sample sizes and the disparity between the number of individuals in each group. Outlier loci were defined as those with an F_{ST} value above the 95th percentile of all F_{ST} values in the empirical dataset.

RESULTS

Sequence and haplotype diversity in China

Measures of nucleotide diversity within the Nanpi population and the Yancheng population were similar. Values of π ranged from 0.003–0.130 nucleotide differences per site (Supplementary Table S2), averaging 0.028 across 40 loci, while haplotype diversity ranged from 0.5 to 1 (within amplicon lengths of 120-660 bp), similar to that observed in Australian populations (Song *et al.*, 2015). Linkage disequilibrium decayed in a manner similar to that of the Australian populations, with the average r^2 (the square of the correlation coefficient between pairs of polymorphic sites) falling below 0.2 for sites separated by 200 bp or more (Supplementary Figure S2).

There was little population structure between Nanpi and Yancheng. F_{ST}, a traditional measure of population structure, was very low (0.03) and an analysis of molecular variance (AMOVA) indicated that less than 1% of the variation in the samples discriminated between the two Chinese populations, despite them being 700km apart. Likewise, analyses using the STRUCTURE program provided little support for population differentiation (Figure 3.2A, Supplementary Table S3). We performed two separate analyses on the combined Nanpi and Yancheng datasets; the 'haplotype' and the 'site' analysis. In the former, haplotypes were determined for each of the 40 loci; many haplotypes were found and all were at low frequencies (many haplotypes occur only once in a population). This approach therefore reduces the power to assign individual moths to a particular genetically defined cluster. By contrast, the 'site' analysis treats sites as independent (i.e. assumes no linkage disequilibrium). As each site is limited to a maximum of 4 states (we have not considered gaps) with most sites being bi-allelic, the estimated allele frequencies resulting from this treatment are substantially higher, increasing the ability of the STRUCTURE algorithm to

strongly assign an individual to one cluster or another. Additionally, the use of polymorphic sites rather than haplotypes increases the number of loci available for analysis by 30-fold (40 loci vs. 1352 sites). The STRUCTURE plot shows that Nanpi and Yancheng are both dominated by a single 'genetically defined population' (red in Figure 3.2A). These results indicate that the geographical separation of the two moth collections is not manifest in the genetic data surveyed here.

Given the lack of evidence for population differentiation, the following analyses in this section were performed by treating Nanpi and Yancheng as a single population. Initially, to address the possibility that diversity could be inflated due to sequencing errors, π and θ (measures of genetic diversity estimated from the average number of pairwise differences and the number of segregating sites, respectively) were estimated using the first 100 bases (hereafter referred to as π_{100} and θ_{100}) from the 5' end (excluding indels) where sequence quality is expected to be superior (Table 3.1). These values were compared to values obtained from the full-length datasets. While there was considerable variance between the two estimates at individual loci, the values averaged across all loci were similar, suggesting increased polymorphism in longer sequences due to sequencing error does not inflate estimates substantially.

As the differences between θ and θ_{100} (and π and π_{100}) were small, Tajima's *D* was estimated using all sites. A negative value was observed at every locus, indicating an excess of rare variants in the combined population. At ten of the 40 loci, these values were significantly different from neutral expectations. Five of the ten are located around the *Cyp303a1* locus, which we have deliberately chosen to sample at a higher density here compared to other loci, as this is a region previously reported to harbour signatures of a selective sweep. Of the remaining five loci with a significantly negative Tajima's *D*, two (BGIBMGA012230 and

BGIBMGA012230F2) are located in the same protein-coding gene (BGIBMGA012230) separated by a distance of 4kb. This gene is predicted to code for a subunit of the CCR4-Not protein complex, a global regulator of gene expression (Collart & Panasenko, 2012). The third locus is in the *Kettin* gene, which codes for a highly-conserved protein involved in insect flight muscle development (Lakey *et al.*, 1993). The fourth, BGIBMGA000615 shares some sequence similarity with CG32030 in *D. melanogaster* (http://flybase.org/) which contains a formin domain. Members of the formin family of proteins have been characterized as playing a role in cytokinesis and cytoskeletal control (Wallar & Alberts, 2003). The fifth, BGIBMGA013328 is a TUDOR-SN protein containing staphylococcal nuclease-like (SN) and Tudor domains. The silkworm TUDOR-SN is thought to be involved in the formation of stress granules (RNA-protein complexes that form when translation initiation is impaired during a stress response) and interacts with components of the RNAi pathway (Zhu *et al.*, 2012; Zhu *et al.*, 2013).

Analysis of intercontinental population structure

To investigate the extent of population structure between *H. armigera* from different continents, eight loci that had been Sanger-sequenced in Australian populations previously (Song *et al*., 2015) were included among the markers sequenced in the Nanpi and Yancheng populations: *Cyp303down1*, *Cyp303down3*, *Cyp305b1*, *Period*, *Phc*, *SCAP*, *Tc* and *Tpi*. An analysis was performed to assess two competing hypotheses, namely, whether the samples from Nanpi, Yancheng and MacIntyre Valley clustered into two or three groups. Replicate runs of the STRUCTURE analysis favoured models involving two genetically defined groups, K=2 (Figure 3.2B, Supplementary Table S4). This result is consistent with the lack of differentiation between the two Chinese populations as seen in the data above, and further suggests that the Chinese samples are distinct from the Australian samples. The differences

between the Australian and Chinese samples are subtle with the two genetically defined groups being reciprocally more abundant in the two countries (Figure 3.2B) and admixed individuals being present in all three populations. If the number of genetic groups is increased to three, then one class (green) is more abundant in Australia. Omitting the *Cyp303a1*-associated loci produced a similar pattern. These patterns should not be over-interpreted however as the lack of distinct clustering could also arise due to a lack of power from an insufficient number of markers and/or missing data (Pritchard *et al.*, 2000).

F_{ST} between the Australian and Chinese samples was generally low (0.09), and an analysis of molecular variance (AMOVA) indicated that 88% of the variation in the samples could be explained by the variation within samples from the same country. The *Cyp303a1*-associated loci exhibited the highest F_{ST} values, with 0.24 and 0.69 at *Cyp303down1* and *Cyp303down3* respectively. F_{ST} at the remaining six loci did not exceed 0.2.

The *Cyp303a1* locus

At seven of the eight loci, diversity in the MacIntyre Valley population did not markedly differ from the Chinese populations, although haplotype diversity appeared slightly elevated in the Australian samples (Table 3.2). Notably, *Cyp303down1* was an exception (the Australian population had about ten fold lower nucleotide diversity than the Chinese populations), consistent with other lines of evidence (extended LD, reduced nucleotide diversity and a skewed frequency spectrum) that support the occurrence of a sweep at or around this locus in Australian populations (Song *et al.*, 2015). We explored the patterns around the *Cyp303a1* locus in the Chinese populations, in particular the frequency of the *Del200* haplotype (the swept allele in Australia which is characterised by an intronic 200bp deletion) and the molecular signatures downstream of the indel polymorphism. Analysis of the amplicon lengths revealed that only 1/94 individuals carried the deletion, and sequencing

confirmed the presence of the *Del200* haplotype in this individual, YC-39 (Figure 3.3). Downstream 1kb of the indel, the most abundant haplotype was one that was shared between Australia and China but approximately 27% of the Chinese individuals had a second haplotype that was not observed in MacIntyre Valley (although the small sample size does not preclude the possibility that this second haplotype is present at a low-to-intermediate frequency in Australia). This locus, *Cyp303down1* also exhibited a pattern of extended LD (r^2>0.2 beyond 400bps) that differed from that of other loci where r^2 typically declines to below 0.2 within 200bps (Supplementary Figure S3).

A similar pattern of differentiation was observed between the Australian (*n*=22) and Chinese (*n*=69) populations in the region 3kb downstream of the *Cyp303* coding sequence (Figure 3.3). The major haplotype in Australia (Haplotype 2) was seen in two individuals from China, one of which was YC-39. As YC-39 was also the sole carrier of the *Del200* allele amongst the 94 Chinese samples in this dataset, we infer that YC-39 carries a haplotype that originated from Australia, and conclude that this locus is highly informative for determining whether an *H. armigera* individual originated from Australia or China. The *Cyp303down3* haplotype is characterised by 14 polymorphic sites, three of which show a distinct difference in allele frequencies between Australia and China (Supplementary Figure S4).

Are there other Z loci that show strong differentiation between Australian and Chinese populations?

The strong discordance between the inter-continental population structure observed at *Cyp303a1* and the other loci surveyed motivated us to extend the study to more loci. We wished to know whether the *Cyp303a1* region was unique – perhaps because of its selective sweep, and whether there were other loci that showed strong inter-continental structure. We therefore turned our attention to the whole-genome sequence dataset generated by Anderson

et al. (2016), and examined the Z chromosome sequences from 19 Australian and four Chinese samples in that dataset. In particular, we measured the weighted F_{ST} (Figure 3.4) across the *H. armigera* Z using sliding windows of 500 SNPs (see Materials and Methods). The plot has two striking features relative to such analyses from other species (e.g. Reinhardt *et al.,* 2014): firstly, the F_{ST} values are distributed over a surprisingly large range (as high as F_{ST}=0.6) and secondly, the extremely high F_{ST} values are clustered along the chromosome so that distinct peaks are observed in the plot. This completely independent analysis is concordant with the EPIC data set to the extent that the *Cyp303a1* locus is highly differentiated between Australia and China and only one of the six loci (the *period* gene) that showed minimal population differentiation in the EPIC data set, shows any Australia-China differentiation in the re-sequencing dataset.

As the whole Z-chromosome dataset had an unbalanced sampling design with four moths sampled from China and 19 from Australia, we used a permutation approach to evaluate the extent to which the empirical dataset deviated from the distribution of F_{ST} values that could be expected given the small sample sizes. Four chromosomes were sampled at random from the combined set of 23 and F_{ST} was calculated. This was repeated 1000 times to give 1000 datasets. Regions in the empirical dataset (represented by black dots in Figure 3.4) that fall outside of a distribution generated by 1000 permutations (grey dots) cannot be attributed to population sampling biases. It is noteworthy that the 95th percentile of the empirical dataset is substantially higher than that of the permuted datasets (represented by black and grey dashed lines in the figure, respectively) supporting the contention of population differentiation between Australia and China.

There are 302 windows that cross a threshold defined by the top 95% most differentiated windows. These loci are therefore of interest as they provide useful markers of population

differentiation. The 302 sliding windows can be grouped into 32 contiguous regions (ranging in size from 9kb to 268kb) and further simplified into seven broad arbitrarily-defined regions based on visual inspection. We used the recently available *H. armigera* genome sequence (Pearce *et al.*, 2017) to identify the loci in these differentiated regions – these genes are listed in Table 3.3 along with credible homology-based functional annotations we can assign to them. Notably, the list includes members of the *ABC* transporter gene family which have been implicated in insecticide resistance (Srinivas *et al.*, 2004; Buss and Callaghan, 2008). Also in the list are *period* (used as an EPIC marker in this study) and *CCR4-Not* which shows a significantly negative Tajima's *D* value in the Chinese populations.

DISCUSSION

Our analysis of the EPIC data found that Australian and Chinese populations of *H. armigera* harboured similar levels of nucleotide diversity. Furthermore, the average pairwise nucleotide divergence between the populations in the EPIC dataset is the same as that calculated in the independent dataset of Anderson *et al.* (2016) where π=0.028. Haplotype diversity generally appears to be slightly elevated in the Australian samples (excluding *Cyp303a1* from the comparison), although there are instances where nucleotide diversity of the Nanpi population exceeds that of MacIntyre Valley. The elevated Australian diversity motivated us to examine all the datasets in the literature to consider whether nucleotide diversity could provide a clue as to where the ancestral *H. armigera* populations arose. Revisiting the mtDNA data in Behere *et al.* (2007) reveals that a Ugandan population of *H. armigera* has higher nucleotide diversity than the Australian population, although haplotype diversity of Australia still exceeds that of Uganda. Anderson *et al.* (2016) found more mitochondrial diversity in Australia but the larger number of Australian samples distorts the interpretation slightly. More data would be required to determine whether there is a geographic region that harbours more diversity than all others.

Another result from our EPIC analyses is the consistent signal of a negative Tajima's *D*, indicating an excess of rare variants in the Chinese populations. This result contrasts with that of Anderson *et al.* (2016) who report that Tajima's *D* was positive in eight of the nine populations they examined, the exception being the Australian population. Even when singleton sites are removed from our analysis, Tajima's *D* remains negative for 29 of the 40 loci, suggesting that our data are robust and sequencing errors do not overly inflate the overall negative pattern. We have two hypotheses to explain the discordance between the datasets. Firstly, all our loci are on the Z-chromosome whereas the data from Anderson *et al.* (2016)

come from across the genome (which consists of 31 chromosomes of roughly similar size). Sex chromosomes are subjected to more efficient selection than autosomes because recessive mutations are exposed in the heterogametic sex (Charlesworth *et al.*, 1987). Furthermore, female heterogamety means that Z-linked loci are subjected to higher mutation rates than autosomal loci because the Z chromosome spends two-thirds of its time in males where spermatogenesis is more mutagenic than oogenesis (Vicoso and Charlesworth, 2006; Sackton *et al.*, 2014). Consequently, the combination of increased efficacy of selection and the increase in mutation rate could produce a relatively greater excess of rare variants on the Z chromosome relative to the autosomes.

The second explanation is that the Tajima's *D* analysis of Anderson *et al.* (2016) is affected by the small sample size of populations (most are between 3 and 5 samples per population). Since Tajima's *D* test is an allele frequency spectrum test, power is much lower with small sample sizes. If we exclude these populations on the basis of small sample size, we are left with the one sample where Anderson *et al.* (2016) have a moderate depth for their Tajima's *D* test, the Australian population (*n*=17) which has a negative Tajima's *D* value. This set of individuals is in fact a subset of the 19 examined here, but is genome-wide rather than limited to the Z-chromsome, as our analysis is. Thus we interpret the negative Tajima's *D* as evidence of a population expansion in the evolutionary history of *H. armigera.* It should be noted that such an expansion would be a much older event and not the recent incursion into the Americas (where insufficient time has passed to leave behind such molecular signatures). It is not clear if this expansion relates to the spread of agriculture or to an even older event such as the availability of new niches during favourable climatic conditions in the Pleistocene (the estimated period of divergence with *Helicoverpa zea*). More data and sophisticated population models would be required to date the expansion.

Our data also speak to the question of whether *H. armigera* from Australia belong to a separate subspecies (Common, 1953). Most of the eight loci where we have EPIC data from China and Australia show no population differentiation between the two countries; the exceptions are the *Cyp303a1*-associated markers. Our previous study (Song *et al.*, 2015) which characterised the *Cyp303a1 Ins200* and *Del200* haplotypes revealed extremely diverged alleles – so diverged that they may be considered 'comet alleles' (Staubach *et al.*, 2012) which may have arisen from introgression with some other unknown species. This suggests two contrasting models to explain the results of the genome-wide STRUCTURE analysis reported by Anderson *et al.* (2016). In one model, the signal for population differentiation comes from a smallish number of highly diverged loci (like *Cyp303a1*) that are distributed patchily across Australian genomes. These may derive from an introgression event in a way analogous to the 'Neanderthal' introgression footprints in non-African modern humans (Green *et al.*, 2010). In the other model, the Australian *H. armigera* differ by small amounts diffused throughout the genome, reflecting variants that accumulated in Australian moths when gene flow to other parts of the world was less than it appears to be now; in other words, a more conventional isolation-by-distance model. The gene flow that established such a difference would have to be less than what is observed currently based on reports such as the repeated incursions of *H. armigera* into South America in recent years.

The F_{ST} scan across the Z chromosome offers some support for the 'Neanderthal' model. There appear to be distinct clusters of elevated F_{ST} arising from a substantially lower baseline. Under an isolation-by-distance model, we may have seen more intermediate values of F_{ST}. These inter-population divergence patterns will be influenced by many factors including recombination rate heterogeneity and the timing and abundance of local selective sweeps Future studies should test other global populations for Z-linked population structure and should contrast these within-species data to between-species divergence with closely related

species.

At a pragmatic level, we have identified multiple loci including *Cyp303a1* as useful markers to identify gene flow from Australia. We believe that the Chinese YC-39 individual, which shows a distinctly Australian haplotype at regions close to *Cyp303a1*, tells us that inter-continental gene flow can occur. Such gene flow would contribute to a general lack of population structure in worldwide populations of *H. armigera,* and the most informative loci to track moth movements will be those that have recently increased in frequency due to strong positive selection locally. Most likely among those will be insecticide resistance genes, which currently include cadherin, ABC transporters, *Cyp337b3*, and *kdr* genes (Martinez-Torres *et al.* 1997; Head *et al.*, 1998; Gahan *et al.*, 2001; Gahan *et al.*, 2010; Joußen *et al.*, 2012). In fact, *Cyp303a1*, which Song *et al.* (2015) found to show the hallmarks of a selective sweep, and which belongs to the cytochrome P450 multigene family that includes classic detoxifying enzymes, may actually be a resistance allele itself. We note that Daly and Fisk (1998) report a Z-linked endosulfan resistance in Australian populations, so *Cyp303a1* may be a candidate gene worth testing for this trait. Similarly, the *ABCB* subfamily of *ABC* transporters is of particular interest as they encompass a class of proteins known as P-glycoproteins which have been shown to be involved in transporting xenobiotics out of the cell (reviewed in Buss and Callaghan, 2008).

Conclusions:

Our study has led us to the following four conclusions. Firstly, specific Z-linked loci (such as *Cyp303a1*) provide useful markers for inter-continental population structure in *Helicoverpa armigera*. Secondly, an Australian haplotype at the *Cyp303a1* gene is found in China and the length of this haplotype indicates that there has been recent gene flow from Australia to China. Thirdly, our deep sampling of Chinese allelic diversity reveals that low frequency,

diverse haplotypes exist at multiple Z-loci within Chinese populations, and even when highly diverged haplotypes are excluded from the analyses, there is a skew towards rare variants in the dataset generally that we interpret as evidence of population expansion. Fourthly, remarkable localized clusters of high F_{ST} values occur across the *H. armigera* Z chromosome, perhaps supporting the proposition that the *Helicoverpa armigera conferta* subspecies may originate from an introgression event from an unknown species.

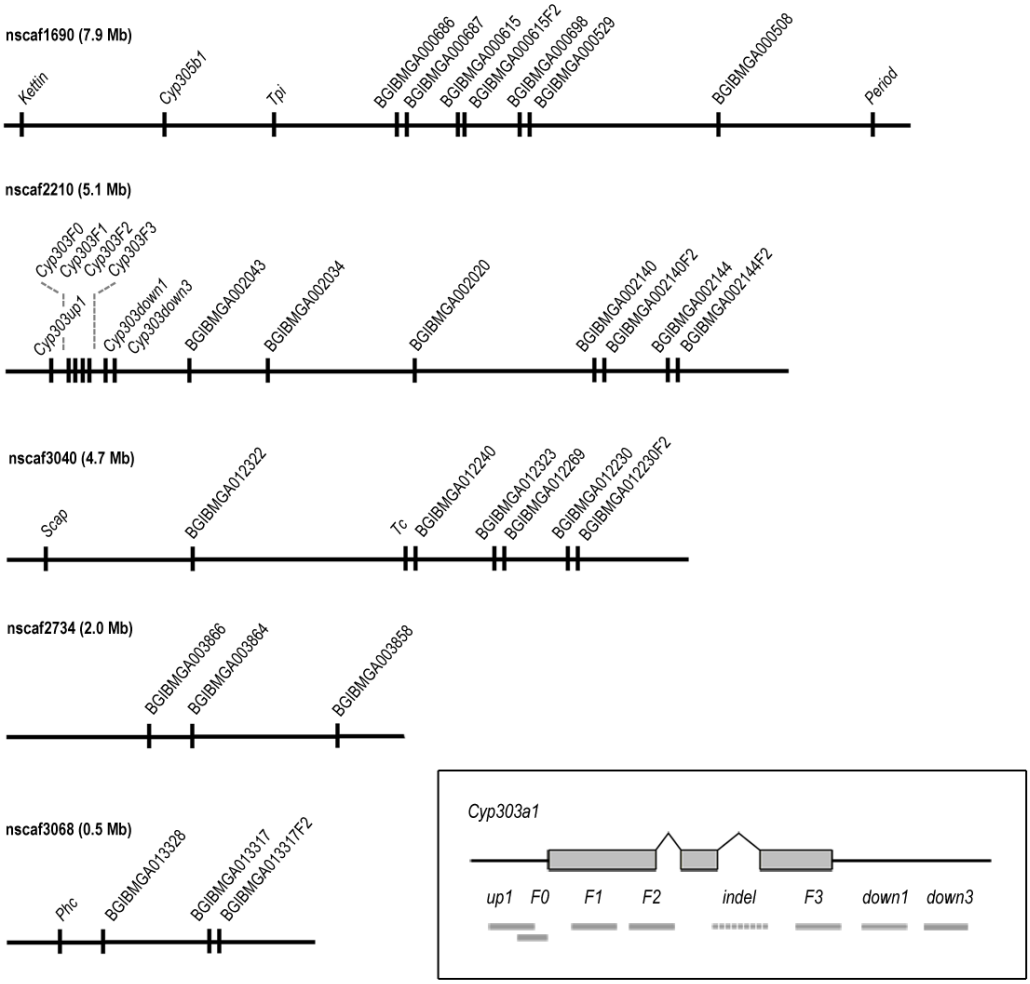


Figure 3.1: Map of EPIC amplicons used in this study. Positions of contigs are based on the five Z chromosome scaffolds of *B. mori*. Numbers in brackets after the scaffold names refer to the lengths of the scaffolds in megabases. Amplicons are named according to the gene they are based on; those with the suffix BGIBMGA refer to *B. mori* orthologs using the names assigned by the silkworm sequencing consortium (*http://silkworm.genomics.org.cn*).

Inset: Amplicon design for the *Cyp303a1* locus.

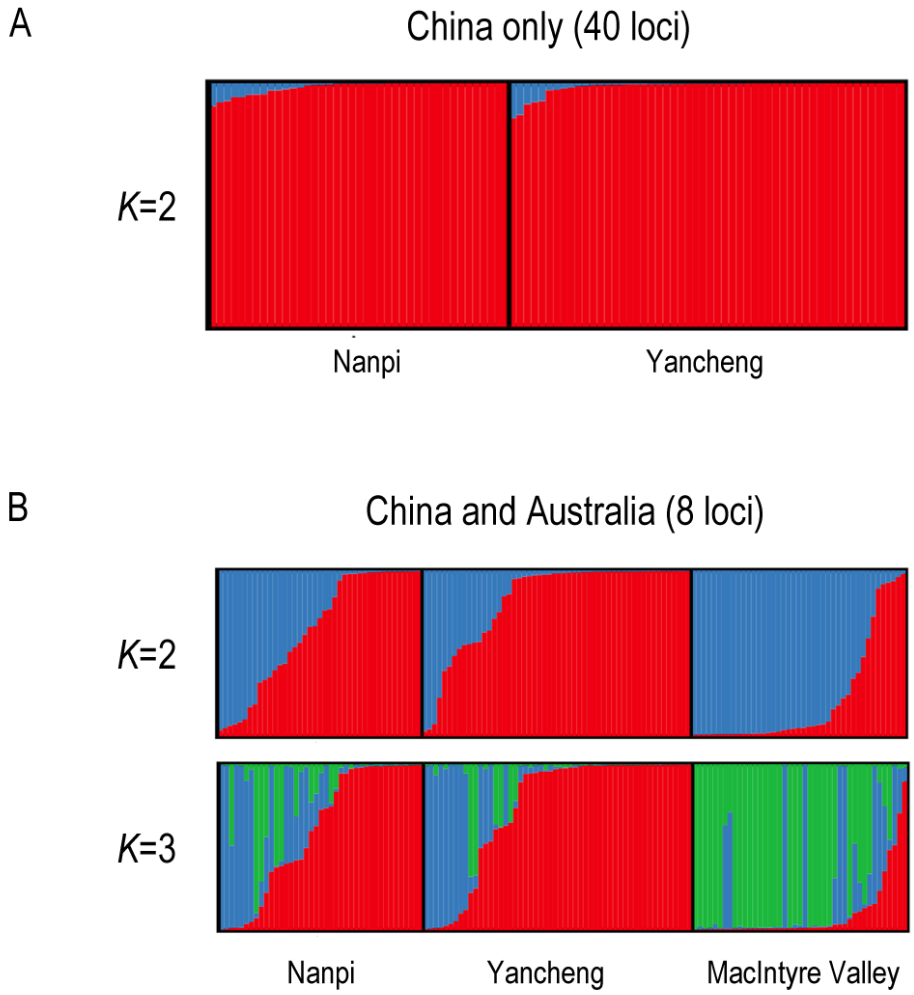


Figure 3.2: A) STRUCTURE plots using 40 loci for the Nanpi and Yancheng populations, with K=2. Each bar (column) represents an individual. Colours and bar heights represent the inferred ancestries of an individual.

B) STRUCTURE plots using 8 loci for the Nanpi, Yancheng and MacIntyre Valley populations. Each bar (column) represents an individual. Colours and bar heights represent the inferred ancestries of an individual, so admixed individuals that have a heritage derived from mixed sources are represented by columns with more than one colour. Only K=2 and K=3 are shown as they represent the values most likely to explain the major structure in the dataset. In the K=2 visualization, MacIntyre Valley has more 'blue' individuals while China has more 'red'. In the K=3 visualization, MacIntyre Valley has more 'green' individuals compared to the primarily red and blue landscape of the Chinese individuals, suggesting that some alleles observed in Australia are rare (or not found) in China. Similarly, the higher incidence of 'red-only' individuals in China suggests that some alleles that are common here are rare in Australia.

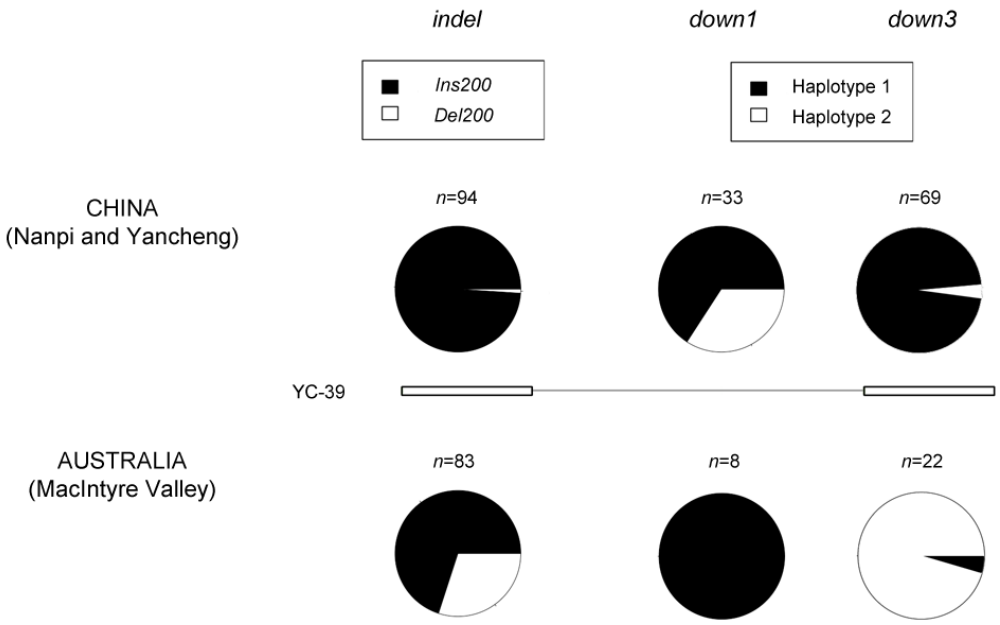


Figure 3.3: Major haplotypes observed in the sequenced regions of the *Cyp303a1* locus in Australia. At the *Cyp303*indel locus, the *Del200* haplotype occurs at approximately 30% frequency in Australia but only appears in one individual, YC-39 from China. Downstream 1kb of the indel, two major haplotypes are observed in China. At the *down3* locus, a different haplotype dominates in the two countries, and individual YC-39 carries a haplotype that is commonly found in Australia. The haplotype at the *down3* locus is primarily defined by two sites, from a total of 14 polymorphic sites at this locus (Supplementary Figure S4).

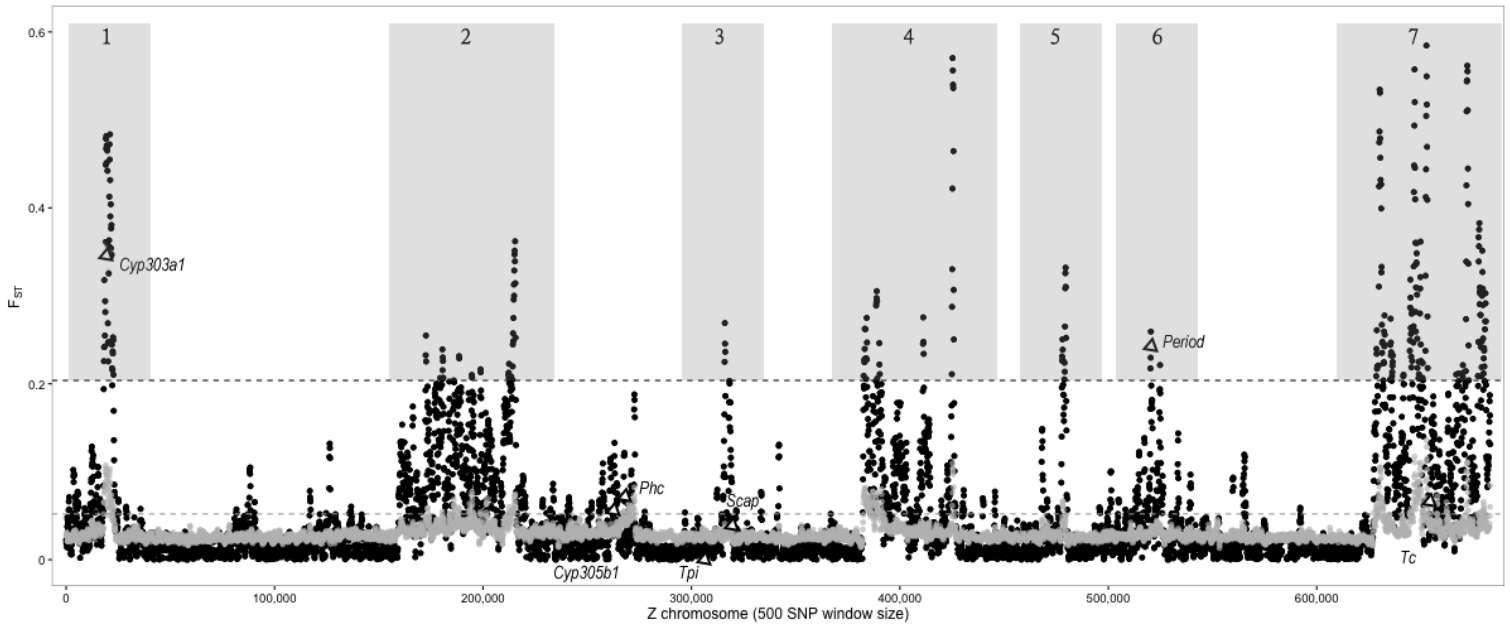


Figure 3.4: Sliding window analysis of weighted F_{ST} across the *H. armigera* Z chromosome for Australian (n=19) and Chinese individuals (n=4). Grey dots represent the 95th percentile of F_{ST} values in each window for 1000 random permutations of the dataset while black dots represent the values from the empirical dataset. The grey dashed line indicates the 95th percentile of the FST values for the permuted datasets while the black dashed line indicates the 95th percentile of the FST values for the empirical dataset. Seven broad visually-defined regions showing high population structure are numbered and indicated by grey shading. The loci that were studied in the initial dataset are shown and their values in this dataset are shown as triangles.

Locus	π	π_{100}	θ	θ_{100}	**Tajima's *D***
BGIBMGA000508 (22)	0.009	0.021	0.018	0.033	-1.69
BGIBMGA000529 (56)	0.082	0.114	0.114	0.144	-0.98
BGIBMGA000615 (68)	0.026	0.016	0.072	0.044	-1.92*
BGIBMGA000615F2 (47)	0.035	0.021	0.069	0.036	-1.49
BGIBMGA000686 (61)	0.030	0.061	0.069	0.113	-1.68
BGIBMGA000687 (66)	0.026	0.023	0.040	0.036	-1.13
BGIBMGA000698 (61)	0.029	0.021	0.079	0.051	-1.70
BGIBMGA002020 (19)	0.037	0.027	0.043	0.034	-0.54
BGIBMGA002034 (54)	0.044	0.031	0.090	0.092	-1.61
BGIBMGA002043 (46)	0.019	0.012	0.029	0.009	-1.22
BGIBMGA002140 (50)	0.032	0.005	0.063	0.018	-1.63
BGIBMGA002140F2 (90)	0.029	0.031	0.050	0.041	-1.37
BGIBMGA002144 (28)	0.048	0.045	0.084	0.081	-1.65
BGIBMGA002144F2 (57)	0.032	0.020	0.066	0.048	-1.48
BGIBMGA003858 (10)	0.009	0.010	0.014	0.011	-1.65
BGIBMGA003864 (52)	0.046	0.040	0.062	0.042	-0.74
BGIBMGA003866 (93)	0.067	0.053	0.092	0.076	-0.91
BGIBMGA012230 (19)	0.004	0.002	0.009	0.006	-2.15**
BGIBMGA012230F2 (83)	0.005	0.002	0.015	0.012	-2.00*
BGIBMGA012240 (88)	0.027	0.007	0.043	0.014	-1.01
BGIBMGA012269 (13)	0.032	0.045	0.048	0.089	-1.00
BGIBMGA012322 (83)	0.027	0.035	0.062	0.058	-1.42
BGIBMGA012323 (31)	0.125	0.095	0.202	0.145	-0.43
BGIBMGA013317 (24)	0.036	0.002	0.056	0.005	-1.24
BGIBMGA013317F2 (84)	0.006	0.006	0.015	0.008	-1.77
BGIBMGA013328 (68)	0.006	0.001	0.011	0.002	-1.86*
Cyp303down1 (33)	0.021	0.032	0.029	0.041	-0.98
Cyp303down3 (69)	0.009	0.004	0.026	0.012	-2.04*
Cyp303F0 (75)	0.013	0.015	0.046	0.043	-2.21*
Cyp303F1 (88)	0.004	0.007	0.015	0.014	-2.30**
Cyp303F2 (83)	0.005	0.004	0.026	0.024	-2.63***
Cyp303F3 (25)	0.016	0.012	0.018	0.021	-0.49
Cyp303up1 (78)	0.023	0.029	0.061	0.057	-1.98*
Cyp305b1 (36)	0.020	0.011	0.027	0.012	-0.94
Kettin (93)	0.007	0.012	0.024	0.049	-2.15*
Period (87)	0.015	0.012	0.032	0.025	-1.34
Phc (59)	0.039	0.018	0.055	0.037	-0.72
Scap (79)	0.034	0.026	0.062	0.085	-1.46

Tc (89)	0.008	0.009	0.014	0.012	-1.37
Tpi (48)	0.048	0.025	0.061	0.036	-0.33
Average	0.028	0.024	0.050	0.043	

Table 3.1: Nucleotide diversity and Tajima's D for all sites, and for the first 100 bases. Figures in brackets after the locus name represent the total number of sequences surveyed.

* $p<0.05$

** $p<0.01$

*** $p<0.001$

Locus	Population	Sample size, n	Haplotype diversity	Nucleotide diversity, π
Cyp303down1	MV	8	0.64	0.002
	Nanpi	15	0.96	0.025
	Yanch.	18	0.93	0.015
Cyp303down3	MV	22	0.65	0.005
	Nanpi	26	0.71	0.012
	Yanch.	44	0.65	0.008
Cyp305b1	MV	21	0.97	0.024
	Nanpi	17	0.95	0.020
	Yanch.	19	0.94	0.021
Period	MV	17	0.97	0.020
	Nanpi	37	0.76	0.021
	Yanch.	51	0.84	0.020
Phc	MV	16	0.95	0.037
	Nanpi	25	0.92	0.042
	Yanch.	34	0.89	0.036
Scap	MV	11	0.95	0.045
	Nanpi	34	0.95	0.036
	Yanch.	46	0.90	0.032
Tc	MV	19	0.99	0.014
	Nanpi	38	0.81	0.009
	Yanch.	52	0.68	0.007
Tpi	MV	19	0.96	0.059
	Nanpi	17	0.95	0.052
	Yanch.	31	0.95	0.044

Table 3.2: Haplotype diversity and nucleotide diversity for Nanpi, Yancheng and MacIntyre Valley at eight loci

HAOGS Reference	Annotation	Region
HaOG205589	BMORI:facilitated trehalose transporter Tret1-like	1
HaOG203061	BMORI:macrophage migration inhibitory factor-like	1
HaOG203059	BMORI:mitochondrial thiamine pyrophosphate carrier-like	1
HaOG216810	BMORI:patj homolog	1
HaOG213506	BMORI:putative zinc finger protein 724-like	1
HaOG203065	BMORI:RNA pseudouridylate synthase domain-containing protein 1-like	1
HaOG214178	BMORI:T-box protein H15-like	1
HaOG203062	BMORI:ubiquitin-conjugating enzyme E2 G2-like	1
HaOG203060	CELEG:P91119 Probable 3',5'-cyclic phosphodiesterase pde-5 BMORI:probable 3',5'-cyclic phosphodiesterase pde-5-like	1
HaOG200007	CYP303A1-Ha	1
HaOG203063	DMELA:A1Z6X0 CG12164	1
HaOG203066	DMELA:Q8SYS6 RE37593p BMORI:alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase B-like	1
HaOG213125	DMELA:Q7Z020 Transient receptor potential cation channel subfamily A member 1 BMORI:transient receptor potential cation channel subfamily A member 1-like	2
HaOG213127	DMELA:Q9VSQ2 CG13310 BMORI:uncharacterized protein LOC101744434	2
HaOG211956	BMORI:tRNA-specific adenosine deaminase 1-like isoform X2	2
HaOG213135	DMELA:A8JUT1 Furrowed, isoform B BMORI:sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1-like	2
HaOG214578	BMORI:rho GTPase-activating protein 8-like	2
HaOG206575	BMORI:titin2	2
HaOG214577	DMELA:Q7YZH1 PHD finger protein rhinoceros BMORI:LOW QUALITY PROTEIN: PHD finger protein rhinoceros-like	2
HaOG214580	DMELA:Q9VS49 CG8600, isoform A	2
HaOG214579	DMELA:Q9W0Q2 Peptidyl-prolyl cis-trans isomerase BMORI:peptidylprolyl isomerase	2
HaOG204759	DMELA:A1ZAJ7 CG15615	3
HaOG216091	BMORI:chromatin-remodeling complex ATPase chain Iswi-like	4
HaOG209512	BMORI:filaggrin-like	4
HaOG216089	BMORI:sodium-dependent noradrenaline transporter	4
HaOG200348	HaABCB1 ALT:HaABC-B-01-1-F	4
HaOG200349	HaABCB2 ALT:HaABC-B-01-2-F	4
HaOG200350	HaABCB3 ALT:HaABC-B-01-3-F	4
HaOG216075	DMELA:A8JMD5 CG34356 BMORI:SCY1-like protein 2-like	4
HaOG216074	BMORI:EF-hand calcium-binding domain-containing protein 1-like	4

HaOG216055	DMELA:P45447 Ecdysone-induced protein 78C BMORI:ecdysone-inducible protein E75-like	4
HaOG209024	BMORI:niemann-Pick C1-like protein 1-like	4
HaOG208377	BMORI:succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial-like	4
HaOG208378	BMORI:succinate dehydrogenase cytochrome b560 subunit, mitochondrial-like	4
HaOG207326	BMORI:insulin-like growth factor 2 mRNA-binding protein 1-like isoform X1	5
HaOG215518	BMORI:pogo transposable element with ZNF domain-like	5
HaOG207329	BMORI:exportin-1-like	5
HaOG212018	BMORI:LIM domain-containing protein jub-like	5
HaOG214354	BMORI:methylcrotonoyl-CoA carboxylase subunit alpha, mitochondrial-like	5
HaOG207328	BMORI:rho guanine nucleotide exchange factor 11-like	5
HaOG207327	DMELA:E1JJM0 FI20063p1	5
HaOG215851	DMELA:P07663 Period circadian protein BMORI:period	6
HaOG203153	DMELA:O76933 Pentaxin-like protein BMORI:uncharacterized protein LOC101736996 isoform X1	7
HaOG203152	DMELA:Q7KW14 Coiled-coil domain-containing protein CG32809 BMORI:coiled-coil domain-containing protein AGAP005037-like	7
HaOG209864	DMELA:Q8MQJ5 CG31122, isoform B	7
HaOG203160	BMORI:protein bric-a-brac 1-like, partial	7
HaOG215757	BMORI:leucine-rich repeat serine/threonine-protein kinase 1-like	7
HaOG203170	BMORI:protein Daple-like	7
HaOG204924	BMORI:protein ELYS-like	7
HaOG213826	BMORI:threonine dehydratase, mitochondrial-like	7
HaOG203168	DMELA:E1JI71 CG32352, isoform E BMORI:biorientation of chromosomes in cell division protein 1-like 1-like	7
HaOG203164	DMELA:O76867 EG:100G10.7 protein BMORI:paraplegin-like	7
HaOG213827	DMELA:Q7KST5 CG8129, isoform A	7
HaOG203169	DMELA:Q9VRT9 CG13293 BMORI:uncharacterized protein LOC101746024	7
HaOG203162	DMELA:Q9W3N7 CG18624, isoform A	7
HaOG203161	HSAPI:A5YKK6 CCR4-NOT transcription complex subunit 1 BMORI:CCR4-NOT transcription complex subunit 1-like	7
HaOG203166	HSAPI:Q13946 High affinity cAMP-specific 3',5'-cyclic phosphodiesterase 7A BMORI:uncharacterized protein LOC101739482	7
HaOG205332	BMORI:Fanconi anemia group M protein-like	7
HaOG205326	BMORI:carbonic anhydrase 1-like	7
HaOG205325	BMORI:H/ACA ribonucleoprotein complex non-core subunit NAF1-like	7

HaOG205324	DMELA:A8JUV8 Terribly reduced optic lobes, isoform G BMORI:basement membrane-specific heparan sulfate proteoglycan core protein-like	7
HaOG205330	DMELA:Q9W4Y2 PDF receptor BMORI:neuropeptide receptor B2	7
HaOG207317	DMELA:P47825 Transcription initiation factor TFIID subunit 4 BMORI:LOW QUALITY PROTEIN: transcription initiation factor TFIID subunit 4-like	7
HaOG207316	DMELA:Q8IQN2 CG5284, isoform B BMORI:H(+)/Cl(-) exchange transporter 3-like	7
HaOG207314	BMORI:LOW QUALITY PROTEIN: adenosine deaminase CECR1-like	7
HaOG207312	BMORI:neuropeptide receptor A18	7
HaOG207310	BMORI:olfactory receptor 60	7
HaOG207311	BMORI:trypsin-1-like	7
HaOG207313	DMELA:Q2HPH2 Peptide receptor GPCR BMORI:neuropeptide receptor A18	7
HaOG207315	DMELA:Q9VT28 Protein furry BMORI:LOW QUALITY PROTEIN: microtubule-associated serine/threonine-protein kinase 2-like	7
HaOG207296	BMORI:beta carbonic anhydrase 1-like	7
HaOG207303	BMORI:G-protein-signaling modulator 2-like	7
HaOG207295	BMORI:mucin-17-like	7
HaOG207299	BMORI:ociad protein isoform 1	7
HaOG207293	BMORI:tetratricopeptide repeat protein 26-like isoform X1	7
HaOG207301	DMELA:O61734 Protein cycle BMORI:Cycle like factor b	7
HaOG207304	DMELA:Q9NFR7 Rapsynoid BMORI:G-protein-signaling modulator 2-like	7
HaOG207300	DMELA:Q9W1X9 OCIA domain-containing protein 1 BMORI:ociad protein isoform 2	7

Table 3.3: Functional annotations for F_{ST} outlier loci identified from the sliding window analysis across the *H. armigera* Z chromosome. HAOG numbers refer to the *H. armigera* official gene set annotated in Pearce *et al.* (2017). Region numbers correspond to those shown in Figure 3.4.

3.3 Supplementary material

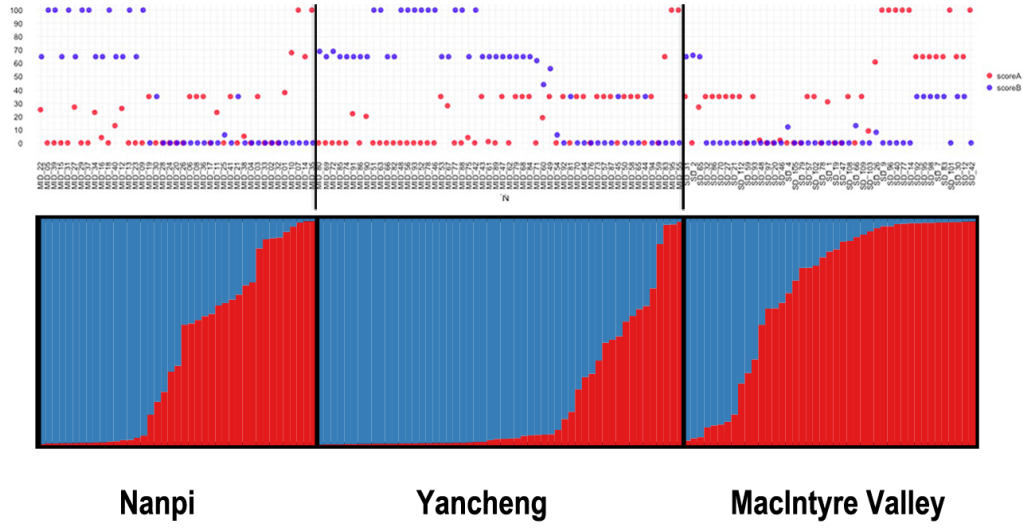


Figure S1: Correlation between a single, randomly-selected graphical representation of a STRUCTURE run and the results of multiple runs. A total of 100 runs were carried out in STRUCTURE using K=2. The dotplot shows the number of times each individual was strongly (probability of 0.95 or higher) assigned to a particular cluster, while the coloured barplot is a typical representation of a STRUCTURE analysis to infer population structure from genotypic data.

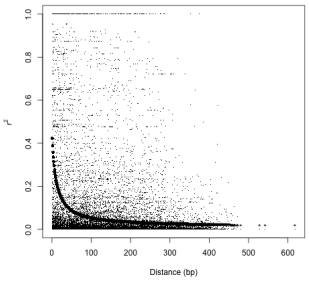


Figure S2: Decay of linkage disequilibrium across 40 loci compared in the Nanpi and Yancheng populations. Data points represent r^2 values from each locus plotted against physical distance in base pairs. The curve shows the decay of LD modelled on the expectations of Hill & Weir (1988).

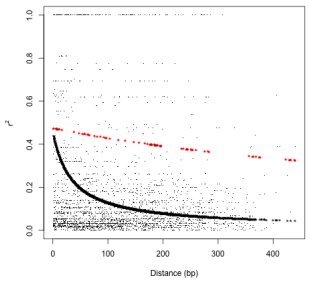


Figure S3: Decay of linkage disequilibrium across eight loci in the MacIntyre Valley, Nanpi and Yancheng populations. Data points represent r^2 values taken from *Cyp303down1*, *Cyp303down3*, *Cyp305b1*, *Period*, *Phc*, *SCAP*, *Tc* and *Tpi*, plotted against physical distance in base pairs. The curve in red models the decay of LD for *Cyp303down1* while the decay of the other seven loci (pooled) is modelled in grey. The models used are described by the formula in Hill & Weir (1988).

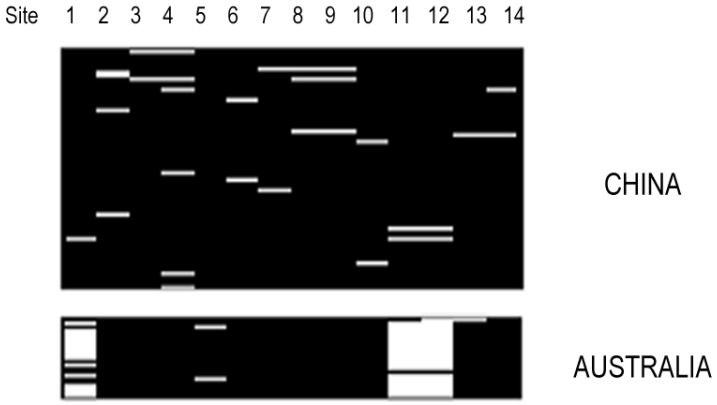


Figure S4: The *Cyp303down3* haplotypes in China (n=69) and Australia (n=22). Each row represents an individual and alleles are arbitrarily coded as black and white. The haplotype is characterised by 14 sites. At three sites (1, 11 and 12), there is a distinct difference in allele frequencies between Chinese and Australian individuals. The major Australian haplotype is primarily defined by Sites 11 and 12.

Primer	Sequence 5'–3'
BGIBMGA000508F	GCCTCCCTCGCGCCAaggttctcttaaacatcaagaaatg
BGIBMGA000508R	GCCTTGCCAGCCCGCttcaccatttaagcttatacaNcg
BGIBMGA000529F	GCCTCCCTCGCGCCAatcgtcaacgccttctactc
BGIBMGA000529R	GCCTTGCCAGCCCGCtctgcgatgttctcgccttg
BGIBMGA000615F	GCCTCCCTCGCGCCAtttggaaggagattcaagagaa
BGIBMGA000615R	GCCTTGCCAGCCCGCccatcttcataatagctgccttt
BGIBMGA000615F2	GCCTCCCTCGCGCCAcgagttcgtcgccaacaa
BGIBMGA000615R2	GCCTTGCCAGCCCGCgaaggagtttgagagcggttt
BGIBMGA000686F	GCCTCCCTCGCGCCAgattgtatggacgacttcagtgc
BGIBMGA000686R	GCCTTGCCAGCCCGCatgtgccgcgagacagtta
BGIBMGA000687F	GCCTCCCTCGCGCCAatggcagctggtatattcatca
BGIBMGA000687R	GCCTTGCCAGCCCGCtcctcgaatatcactccacaca
BGIBMGA000698F	GCCTCCCTCGCGCCAgacgaagagagccttttatttca
BGIBMGA000698R	GCCTTGCCAGCCCGCgcattgctcattttgaatgg
BGIBMGA002020F	GCCTCCCTCGCGCCAcatttccagatgttcaactacca
BGIBMGA002020R	GCCTTGCCAGCCCGCgcgcaaagtcgggtaaatag
BGIBMGA002034F	GCCTCCCTCGCGCCAcgcttgaaggttctccatga
BGIBMGA002034R	GCCTTGCCAGCCCGCcgcgtgggtatctccactt
BGIBMGA002043F	GCCTCCCTCGCGCCAgtgccgtaccacaacttcg
BGIBMGA002043R	GCCTTGCCAGCCCGCatggtgacNgttatgttgaagtga
BGIBMGA002140F	GCCTCCCTCGCGCCAcgttcaaccgaactatttgc
BGIBMGA002140R	GCCTTGCCAGCCCGCaccaagttgcccactatcac
BGIBMGA002140F2	GCCTCCCTCGCGCCAgggtacagtcaccgcacaa
BGIBMGA002140R2	GCCTTGCCAGCCCGCacgacgacgcctactatcag
BGIBMGA002144F	GCCTCCCTCGCGCCAccggataaacgtcagtaaacag
BGIBMGA002144R	GCCTTGCCAGCCCGCttgcatacgtcttcacacactg
BGIBMGA002144F2	GCCTCCCTCGCGCCAgcgaaagataacgccaagttc
BGIBMGA002144R2	GCCTTGCCAGCCCGCcttctcccagcaacagacc
BGIBMGA003858F	GCCTCCCTCGCGCCAagatacgggacacgtttacgc
BGIBMGA003858R	GCCTTGCCAGCCCGCcattggtggtcggcttctac
BGIBMGA003864F	GCCTCCCTCGCGCCAggtggtatcaatggcgatg
BGIBMGA003864R	GCCTTGCCAGCCCGCttgtcggccatgtcgttc
BGIBMGA003866F	GCCTCCCTCGCGCCAcgcacagttctcacaggaaa
BGIBMGA003866R	GCCTTGCCAGCCCGCgtggtgtcagggtcgaactc

BGIBMGA012230F	GCCTCCCTCGCGCCAcacagttcaagaaggatttgga
BGIBMGA012230R	GCCTTGCCAGCCCGCttgatgagttcgatgaaggtg
BGIBMGA012230F2	GCCTCCCTCGCGCCAaggagctgcgctacaacatc
BGIBMGA012230R2	GCCTTGCCAGCCCGCccattcgcggaggagatt
BGIBMGA012240F	GCCTCCCTCGCGCCAagttgcgagagttgccagt
BGIBMGA012240R	GCCTTGCCAGCCCGctgatcgtgcggcattcat
BGIBMGA012269F	GCCTCCCTCGCGCCAttcggactgtgtatgcctca
BGIBMGA012269R	GCCTTGCCAGCCCGCgcagtcgcgttgaagtcg
BGIBMGA012322F	GCCTCCCTCGCGCCAgtgttggtcacatacctcatgc
BGIBMGA012322R	GCCTTGCCAGCCCGCcaatcggctgcggtcttc
BGIBMGA012323F	GCCTCCCTCGCGCCAaggaacagccgatggatatg
BGIBMGA012323R	GCCTTGCCAGCCCGCcaggaatgggtctcggta
BGIBMGA013317F	GCCTCCCTCGCGCCAtcaagagcgaatggatttgt
BGIBMGA013317R	GCCTTGCCAGCCCGCtttgctggagacgtggaaa
BGIBMGA013317F2	GCCTCCCTCGCGCCAttccttcgcgatcactcac
BGIBMGA013317R2	GCCTTGCCAGCCCGCgacctgggtacacgccatt
BGIBMGA013328F	GCCTCCCTCGCGCCAagacaagatggagaatcagagc
BGIBMGA013328R	GCCTTGCCAGCCCGCgcattgcttacgtagttcgtc
*Cyp303down1*F	GCCTCCCTCGCGCCAgcactccacagcggttattt
*Cyp303down1*R	GCCTTGCCAGCCCGctgatagaaatccaaaggaagtca
*Cyp303down3*F	GCCTCCCTCGCGCCActaattgaccgacatgctaggg
*Cyp303down3*R	GCCTTGCCAGCCCGCtaatcactgggttgcttctgg
*Cyp303*F0	GCCTCCCTCGCGCCAgggactttgctatgcgaaac
*Cyp303*R0	GCCTTGCCAGCCCGCtgtatgacggctgcatcg
*Cyp303*F1	GCCTCCCTCGCGCCAggagaacgccgaggagtatt
*Cyp303*R1	GCCTTGCCAGCCCGCacttcctgctccctcaccag
*Cyp303*F2	GCCTCCCTCGCGCCAcggcacaaagagaggtttga
*Cyp303*R2	GCCTTGCCAGCCCGCttcttacctctggtataaaatatcc
*Cyp303*F3	GCCTCCCTCGCGCCAtgcgtacgacaattgaataaata
*Cyp303*R3	GCCTTGCCAGCCCGCgtgtaaattcgatgctaacaaga
*Cyp303up1*F	GCCTCCCTCGCGCCAcgtcgcctgaacttacaaca
*Cyp303up1*R	GCCTTGCCAGCCCGCtcttcggcttatgggtatcg
*Cyp305b1*F	GCCTCCCTCGCGCCAacacgtctgcgtttctccaa
*Cyp305b1*R	GCCTTGCCAGCCCGCggtgtagccaatataccaatcaac
*Kettin*F	GCCTCCCTCGCGCCAggctcatcaatggcaacac
*Kettin*R	GCCTTGCCAGCCCGCggcctgtggtcaataactcc

*Period*F	GCCTCCCTCGCGCCAatggcaatgggcagcaac
*Period*R	GCCTTGCCAGCCCGCgcactggttggatggtagga
*Phc*F	GCCTCCCTCGCGCCAtacgcgaagatgtggtacaagg
*Phc*R	GCCTTGCCAGCCCGCgcgcacgttgttgataatcc
*Scap*F	GCCTCCCTCGCGCCAtggactggtgtcaggcttat
*Scap*R	GCCTTGCCAGCCCGCcgggtacatttccttcagctt
*Tc*F	GCCTCCCTCGCGCCAaaatgtgtgctcaagagttgg
*Tc*R	GCCTTGCCAGCCCGCttgttgatagcttcgcaagagt
*Tpi*F	GCCTCCCTCGCGCCAattcgttgttggtggtaactgga
*Tpi*R	GCCTTGCCAGCCCGCtaccgatagcccaaactggt

Table S1: List of 454 primer sequences. The first 15 bases (uppercase) are the universal tails, with a different sequence for forward and reverse primers. The remaining bases (lowercase) are the locus-specific sequence.

Locus	**n**	**Number of sites (bp)**[a]	**S**[b]		**Haplotype diversity**	π	
BGIBMGA000508 **(22)**	Nanpi (13)	472	21	29	0.974	0.010	0.009
	Yanch. (9)		13			0.009	
BGIBMGA000529 **(56)**	Nanpi (24)	312–367	122	115	0.999	0.087	0.082
	Yanch. (32)		123			0.078	
BGIBMGA000615 **(68)**	Nanpi (26)	179–188	38	52	0.788	0.043	0.026
	Yanch. (42)		37			0.025	
BGIBMGA000615F2 **(47)**	Nanpi (20)	347–385	99	92	0.952	0.038	0.035
	Yanch. (27)		64			0.032	
BGIBMGA000686 **(61)**	Nanpi (29)	392–419	92	105	0.980	0.030	0.030
	Yanch. (32)		79			0.029	
BGIBMGA000687 **(66)**	Nanpi (30)	224–251	45	32	0.989	0.027	0.026
	Yanch. (36)		27			0.026	
BGIBMGA000698 **(61)**	Nanpi (23)	130–156	40	37	0.950	0.031	0.029
	Yanch. (38)		28			0.029	
BGIBMGA002020 **(19)**	Nanpi (8)	188	16	26	0.994	0.034	0.037
	Yanch. (11)		23			0.042	
BGIBMGA002034 **(54)**	Nanpi (22)	282–325	79	104	0.859	0.046	0.044
	Yanch. (32)		92			0.044	
BGIBMGA002043 **(46)**	Nanpi (26)	411–432	47	52	0.976	0.021	0.019
	Yanch. (20)		25			0.016	

BGIBMGA002140 **(50)**	Nanpi (13)	388–391	42	98	0.979	0.029	0.032
	Yanch. (37)		90			0.035	
BGIBMGA002140F2 **(90)**	Nanpi (37)	424–425	68	87	0.975	0.028	0.029
	Yanch. (53)		73			0.029	
BGIBMGA002144 **(28)**	Nanpi (10)	343–373	65	84	0.988	0.048	0.048
	Yanch. (18)		78			0.049	
BGIBMGA002144F2 **(57)**	Nanpi (24)	375–399	63	95	0.978	0.030	0.032
	Yanch. (33)		90			0.033	
BGIBMGA003858 **(10)**	Nanpi (8)	446–487	16	18	0.867	0.010	0.009
	Yanch. (2)		3			0.006	
BGIBMGA003864 **(52)**	Nanpi (19)	223–247	53	56	0.878	0.052	0.046
	Yanch. (33)		51			0.044	
BGIBMGA003866 **(93)**	Nanpi (40)	229–241	61	87	0.881	0.063	0.067
	Yanch. (53)		81			0.073	
BGIBMGA012230 **(19)**	Nanpi (13)	443	7	13	0.868	0.003	0.004
	Yanch. (6)		7			0.006	
BGIBMGA012230F2 **(83)**	Nanpi (30)	441–453	14	31	0.752	0.004	0.005
	Yanch. (53)		27			0.006	
BGIBMGA012240 **(88)**	Nanpi (38)	274–287	34	55	0.888	0.027	0.027
	Yanch. (50)		50			0.027	
BGIBMGA012269 **(13)**	Nanpi (4)	356–366	24	49	1.000	0.034	0.032
	Yanch. (9)		42			0.031	
BGIBMGA012322 **(83)**	Nanpi (38)	241–249	38	58	0.994	0.023	0.027

	Yanch. (45)		53			0.031	
BGIBMGA012323 **(31)**	Nanpi (8)	204–295	93	115	0.987	0.111	0.125
	Yanch. (23)		113			0.130	
BGIBMGA013317 **(24)**	Nanpi (14)	356–374	35	69	0.867	0.027	0.036
	Yanch. (10)		56			0.046	
BGIBMGA013317F2 **(84)**	Nanpi (39)	289–300	13	20	0.637	0.006	0.006
	Yanch. (45)		16			0.006	
BGIBMGA013328 **(68)**	Nanpi (28)	302	12	14	0.772	0.006	0.006
	Yanch. (40)		11			0.006	
Cyp303down1 **(33)**	Nanpi (15)	472–521	41	50	0.991	0.023	0.021
	Yanch. (18)		35			0.016	
Cyp303down3 **(69)**	Nanpi (26)	446–455	33	51	0.926	0.012	0.009
	Yanch. (43)		31			0.008	
Cyp303F0 **(75)**	Nanpi (34)	125–133	14	26	0.757	0.014	0.013
	Yanch. (41)		18			0.012	
Cyp303F1 **(88)**	Nanpi (38)	660	18	49	0.882	0.003	0.004
	Yanch. (50)		36			0.004	
Cyp303F2 **(83)**	Nanpi (35)	216	14	25	0.503	0.004	0.005
	Yanch. (48)		19			0.005	
Cyp303F3 **(25)**	Nanpi (4)	439–463	19	30	0.980	0.023	0.016
	Yanch. (21)		27			0.014	
Cyp303up1 **(78)**	Nanpi (34)	284–377	59	79	0.980	0.022	0.023
	Yanch. (44)		67			0.024	

Cyp305b1 **(36)**	Nanpi (17)	460–470	35	48	0.940	0.020	0.020
	Yanch. (19)		36			0.021	
Kettin **(93)**	Nanpi (41)	446–447	33	48	0.931	0.007	0.007
	Yanch. (52)		34			0.007	
Period **(87)**	Nanpi (37)	248–280	27	35	0.815	0.010	0.016
	Yanch. (50)		34			0.019	
Phc **(59)**	Nanpi (25)	286–394	77	66	0.906	0.046	0.039
	Yanch. (34)		58			0.037	
Scap **(79)**	Nanpi (34)	419–446	86	111	0.929	0.037	0.034
	Yanch. (45)		74			0.032	
Tc **(89)**	Nanpi (38)	449	20	31	0.738	0.009	0.008
	Yanch. (51)		22			0.007	
Tpi **(48)**	Nanpi (17)	463	78	109	0.950	0.052	0.048
	Yanch. (31)		91			0.045	

Table S2: Nucleotide diversity and haplotype diversity for 40 loci surveyed in this study. Figures in brackets after the locus name represent the total number of sequences surveyed. Where figures are presented in two columns under a single heading, the left column reports the estimates for an individual population while the right column reports the estimates after pooling the sequences from all populations.

* $p<0.05$

** $p<0.01$

*** $p<0.001$

[a] *the number of sites is presented as a range due to the differing subsets of indel polymorphisms present in different populations. Since gapped sites are excluded from this analysis, the lower boundary represents the number of sites considered when alleles from both populations are pooled.*

[b] *number of segregating sites including singletons.*

Inferred ancestry of individuals:

Label	(% missing)	Pop	Inferred clusters	
MID-01	(52)	1	0.998	0.002
MID-02	(72)	1	0.906	0.094
MID-03	(43)	1	0.999	0.001
MID-04	(29)	1	0.955	0.045
MID-05	(25)	1	1.000	0.000
MID-06	(59)	1	1.000	0.000
MID-07	(53)	1	0.921	0.079
MID-08	(58)	1	0.998	0.002
MID-09	(38)	1	0.999	0.001
MID-10	(67)	1	0.999	0.001
MID-11	(29)	1	0.998	0.002
MID-12	(22)	1	0.999	0.001
MID-13	(40)	1	0.975	0.025
MID-14	(40)	1	0.952	0.048
MID-15	(20)	1	0.999	0.001
MID-16	(29)	1	1.000	0.000
MID-17	(33)	1	0.999	0.001
MID-18	(34)	1	0.998	0.002
MID-19	(32)	1	0.998	0.002
MID-20	(49)	1	0.954	0.046
MID-21	(57)	1	0.943	0.057
MID-22	(58)	1	0.999	0.001
MID-23	(41)	1	0.997	0.003
MID-24	(53)	1	0.998	0.002
MID-25	(24)	1	0.927	0.073
MID-26	(35)	1	0.993	0.007
MID-27	(27)	1	0.970	0.030
MID-28	(47)	1	0.991	0.009
MID-29	(40)	1	1.000	0.000
MID-30	(33)	1	1.000	0.000
MID-31	(25)	1	0.943	0.057
MID-32	(62)	1	0.999	0.001
MID-33	(31)	1	0.994	0.006
MID-34	(27)	1	1.000	0.000

MID-35	(33)	1	0.999	0.001
MID-36	(27)	1	0.993	0.007
MID-37	(65)	1	1.000	0.000
MID-38	(45)	1	0.971	0.029
MID-39	(41)	1	0.980	0.020
MID-40	(37)	1	1.000	0.000
MID-41	(16)	1	0.982	0.018
MID-42	(48)	2	0.998	0.002
MID-43	(29)	2	0.994	0.006
MID-44	(37)	2	0.993	0.007
MID-45	(41)	2	0.972	0.028
MID-46	(51)	2	0.978	0.022
MID-47	(37)	2	0.914	0.086
MID-48	(42)	2	0.999	0.001
MID-49	(36)	2	0.993	0.007
MID-50	(36)	2	0.973	0.027
MID-51	(35)	2	0.999	0.001
MID-52	(30)	2	0.988	0.012
MID-53	(51)	2	0.998	0.002
MID-54	(47)	2	0.996	0.004
MID-55	(23)	2	0.998	0.002
MID-56	(43)	2	0.998	0.002
MID-57	(14)	2	0.998	0.002
MID-58	(28)	2	0.927	0.073
MID-59	(38)	2	0.994	0.006
MID-60	(32)	2	0.999	0.001
MID-61	(38)	2	0.983	0.017
MID-62	(24)	2	1.000	0.000
MID-63	(30)	2	1.000	0.000
MID-64	(24)	2	0.997	0.003
MID-65	(42)	2	0.997	0.003
MID-66	(27)	2	1.000	0.000
MID-67	(35)	2	0.995	0.005
MID-68	(38)	2	0.998	0.002
MID-69	(46)	2	0.999	0.001
MID-70	(29)	2	0.922	0.078

MID-71	(28)	2	0.999	0.001
MID-72	(36)	2	1.000	0.000
MID-73	(28)	2	0.999	0.001
MID-74	(38)	2	1.000	0.000
MID-75	(20)	2	0.996	0.004
MID-76	(27)	2	0.999	0.001
MID-77	(41)	2	1.000	0.000
MID-78	(15)	2	1.000	0.000
MID-79	(62)	2	0.998	0.002
MID-80	(31)	2	0.999	0.001
MID-81	(26)	2	0.995	0.005
MID-82	(17)	2	0.999	0.001
MID-83	(25)	2	0.995	0.005
MID-84	(33)	2	0.996	0.004
MID-85	(28)	2	1.000	0.000
MID-86	(13)	2	1.000	0.000
MID-87	(23)	2	0.989	0.011
MID-88	(52)	2	0.992	0.008
MID-89	(71)	2	0.998	0.002
MID-90	(37)	2	0.999	0.001
MID-91	(40)	2	1.000	0.000
MID-92	(36)	2	0.999	0.001
MID-93	(25)	2	1.000	0.000
MID-94	(42)	2	0.870	0.130
MID-95	(59)	2	0.856	0.144

Table S3: Output of STRUCTURE analysis used to generate Figure 2A showing inferred ancestry of individuals. Column names refer to individual names (labels), percentage of missing data for each individual, population to which each individual belongs (1=Nanpi, 2=Yancheng) and inferred clusters. The analysis was run using K=2 for a total of 1352 sites under a model incorporating admixture and independent allele frequencies between populations, without using prior population information.

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln"(K)|	ΔK
1	10	-8637.08	0.59	—	—	—
2	10	-7835.75	23.65	801.33	579.68	24.51
3	10	-7614.10	148.64	221.65	141.08	0.95
4	10	-7533.53	71.42	80.57	533.88	7.48
5	10	-7986.84	1391.28	-453.31	655.73	0.47
6	10	-7784.42	915.85	202.42	581.56	0.63
7	10	-8163.56	1558.53	-379.14	887.71	0.57
8	10	-7654.99	310.14	508.57	—	—

Table S4: Results of the Evanno method (Evanno *et al.*, 2005) for evaluating ΔK for the Nanpi, Yancheng and MacIntyre Valley populations using 310 sites. The most likely value of K (number of genetically defined populations) was evaluated as that which corresponds to the largest value of ΔK.

REFERENCES

Evanno, G., **Regnaut, S.** & **Goudet, J.** (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**, 2611–2620, 2611–2620. ISSN 1365-294X. doi:10.1111/j.1365-294X.2005.02553.x.

Hill, W.G. & **Weir, B.S.** (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33**, 54–78, 54–78. ISSN 0040-5809. doi:10.1016/0040-5809(88)90004-4.

Chapter 4

Transcriptome analyses of the induction and selection response to fenvalerate in *Helicoverpa armigera*

4.1 Introduction

As a polyphagous pest of economically important crops such as cotton, corn and soybean, *H. armigera* is subjected to several classes of insecticides in the field. Fenvalerate, a synthetic pyrethroid, has been used for control since the 1970s, and resistance in Australian populations of *H. armigera* was detected in the early 1980s (Gunning *et al.*, 1984). Some mechanisms of resistance identified were target-site insensitivity, metabolic detoxification and reduced penetration through the cuticle (Gunning *et al.*, 1991). The introduction of an integrated pest management strategy that resulted in decreased pyrethroid use appeared to change the frequencies of resistance mechanisms whereby resistance due to target-site insensitivity declined while factors associated with detoxification and reduced penetration through the cuticle increased (Gunning *et al.*, 1991; Forrester *et al.*, 1993). Resistance was often found to be suppressed by piperonyl butoxide (PBO), suggesting that detoxification was mediated by CYPs that catalyse monooxygenase reactions, although PBO may also act to increase penetration of the insecticide and/or as a synergist to suppress esterase activity (Kennaugh *et al.*, 1993; Gunning *et al.*, 1995; Young *et al.*, 2005).

4.1.1 Cytochrome P450s

The CYPs are a large and diverse family of proteins found in all eukaryotes and some prokaryotes. They have broad substrate specificity and are known to catalyse at least 60 chemically distinct reactions (Feyereisen, 2005). A CYP enzyme is composed of a single polypeptide of about 500 amino acids, encoded for by a single gene. The active site contains a heme iron centre and most CYPs require a redox partner to reduce the iron in catalytic reactions. In spite of great sequence diversity, the basic fold of the protein is highly conserved (Mestres, 2005; Poulos and Johnson, 2005). Sizes of the CYP gene family differ across insect species, and evidence suggests that species with a broad host range have greater numbers of CYPs compared to specialist feeders as they are exposed to a broader range of substrates including pesticides and plant xenobiotics (Good *et al.*, 2014; Rane *et al.*, 2016). In addition to their roles in detoxification and metabolism of

xenobiotics, CYPs also have important roles in insect growth and developmental pathways (Feyereisen, 1999, 2005).

In studies of *H. armigera* from around the world, several CYPs have been implicated in pyrethroid resistance. Members of the CYP6B and CYP9A subfamilies are especially prominent, although their contribution to the resistance phenotype are not well understood (Ranasinghe *et al.*, 1998; Yang *et al.*, 2004; Grubor and Heckel, 2007; Zhang *et al.*, 2010; Zhou *et al.*, 2010b; Xu *et al.*, 2016). In most cases, a causal link between overexpression of CYPs and insecticide metabolism has not been well established. A notable exception is CYP337B3, which has been demonstrated to metabolise fenvalerate into 4'-hydroxyfenvalerate, a non-toxic compound, in a *H. armigera*-derived cell line (Joußen *et al.*, 2012). Subsequently, CYP6B6 was also found to metabolise esfenvalerate into 4'-hydroxyfenvalerate when co-expressed with NADPH cytochrome P450 reductase (HaCPR) in *E. coli* cells (Tian *et al.*, 2017). CYP9A12 and CYP9A14 have been shown to metabolise esfenvalerate in yeast cells but their metabolites have not been identified (Yang *et al.*, 2008). Members of the CYP4 subfamily have also been found to be overexpressed in pyrethroid-resistant strains of *H. armigera* in the field (Pittendrigh *et al.*, 1997; Brun-Barale *et al.*, 2010) but their metabolic capabilities have not been fully elucidated. Interestingly, CYP4G1 has a role in cuticle formation as it is capable of producing hydrocarbons from aldehydes in *D. melanogaster*, functioning as an oxidative decarbonylase in the presence of a cytochrome P450 reductase (Qiu *et al.*, 2012). Flies deficient in CYP4G1 showed high mortality upon adult emergence and were highly susceptible to desiccation.

4.1.2 CYP337B3

CYP337B3 provides an unusual example of how resistance mechanisms could arise. The gene was identified through linkage mapping involving an Australian fenvalerate-resistant strain crossed to a susceptible strain (Heckel *et al.*, 1998; Wee *et al.*, 2008). The study employed a DNA-based genotyping assay to distinguish between resistant and susceptible progeny of a mapping cross, and a cDNA-AFLP (amplified-fragment length polymor-

phism) assay to identify constitutive expression differences between unexposed cohorts of resistant and susceptible individuals. Of the 525 transcript-derived fragments found to be differentially expressed between the resistant and susceptible pools, two CYPs were found, of which one was *Cyp337b1*, a progenitor of *Cyp337b3*. The *Cyp337b3* gene is a chimera that arose from an unequal crossing-over event between its parental genes, *Cyp337b1* and *Cyp337b2* which are adjacent on chromosome 15 (Joußen *et al.*, 2012). The product of *Cyp337b3* is almost identical to *Cyp337b2* in its first 177 amino acids and *Cyp337b1* in the last 315 amino acids. Neither of the parental genes (nor their allelic variants which differed by up to 29 synonymous and 8 non-synonymous nucleotide substitutions) resulted in products that could metabolise fenvalerate although their enzymatic abilities for other substrates were retained. *Cyp337b3* is thus a functional chimera of two functional genes. Interestingly, *Cyp337b3* is allelic to both its progenitor genes in that homozygous individuals either carry two copies of *Cyp337b3* or two copies of *Cyp337b1* and *Cyp337b2* which are inherited together.

To demonstrate the causal link between CYP337B3 and fenvalerate resistance *in vivo*, Joußen *et al.* (2012) generated two lines from a strain which was heterozygous for *Cyp337b1*, *Cyp337b2* and *Cyp337b3*. One strain was bred to carry only *Cyp337b1* and *Cyp337b2* while the other carried only *Cyp337b3*. Homozygotes from both lines as well as heterozygotes resulting from crosses between the two lines were challenged with topical applications of fenvalerate. The *Cyp337b3* homozygotes were found to be over 40-fold more resistant than *Cyp337b1–Cyp337b2* homozygotes, whereas the heterozygotes exhibited an intermediate level of resistance, with a resistance factor of 12.4. Topical application of the metabolite, 4'-hydroxyfenvalerate onto *Cyp337b1–Cyp337b2* homozygotes (the susceptible strain) was also found to be non-toxic, providing further evidence for the role of CYP337B3 in detoxifying fenvalerate *in vivo*.

A study of a cypermethrin-resistant strain of *H. armigera* from Pakistan revealed the presence of a new *Cyp337b3* allele which resulted in a product that was capable of metabolising cypermethrin to 4'-hydroxycypermethrin (Rasool *et al.*, 2014). In this *Cyp337b3v2* allele, the intron length and sequence as well as the position of the crossing-over event

differ from that of the allele identified by Joußen *et al.* (2012), leading the authors to speculate that *Cyp337b3v2* arose independently. However, the parental *Cyp337b1* and *Cyp337b2* genes have not been observed in Pakistani populations, which casts some doubt on the parsimony of this hypothesis. The *Cyp337b3v2* allele was also found to be at high frequency (96–100%) in field populations of fenvalerate-resistant Chinese *H. armigera*, yet resistance levels were not correlated with allele frequency in laboratory-maintained populations where frequencies of *Cyp337b3* were allowed to drop to between 21% and 73% (Han *et al.*, 2015). Nevertheless, *Cyp337b3* remained absent from the susceptible population, which supports the role of CYP337B3 in fenvalerate resistance although it is likely that other CYPs contribute to the resistance phenotype.

4.1.3 Study aims and hypotheses

Insects have several means by which they can respond to the threats posed by xenobiotics. One strategy is to respond only when the substrate is encountered. Compared to constitutively expressing an arsenal of enzymes, this induction response is presumably less taxing on the insect as it is only required to mobilise its resources at the time the threat is encountered. By studying the induction response to a particular insecticide, we may be able to identify candidate genes that contribute to the resistance phenotype, particularly genes that are involved in metabolic detoxification (Willoughby *et al.*, 2006). This can be complemented by studying the constitutive differences between a selected (resistant) and unselected (susceptible) strain, whereby genes that contribute to the resistance phenotype through mechanisms other than metabolic detoxification may also be identified.

In this study, a comparative analysis was performed between an unselected strain and a strain derived from the same genetic background but selected for resistance to fenvalerate. The strains were exposed to a low, non-lethal dose of fenvalerate then sampled at four different timepoints over a 24-hour period. RNA-Seq was performed to identify candidate genes involved in the transcriptional response of *H. armigera* to fenvalerate. Given the substantial number of studies implicating CYPs in pyrethroid resistance, an *a priori* expectation is that members of this gene family will be enriched in the set of

differentially-expressed transcripts. Several other gene families have also been associated with xenobiotic metabolism (Li *et al.*, 2007; Labbé *et al.*, 2011; Ahn *et al.*, 2012; Dermauw and Van Leeuwen, 2014; Li *et al.*, 2016), so we hypothesise that in the set of differentially-expressed transcripts between the exposed and unexposed cohorts (the induction response) as well as between the selected and unselected strains (the selection response), gene families involved in detoxification such as the GSTs, CCEs, ABCs and UGTs will also be over-represented.

We also aim to provide a preliminary look at candidates for resistance beyond the CYPs, GSTs, CCEs, ABCs and UGTs, as xenobiotic metabolism will not be confined to these five families. Further, metabolic detoxification is only one of several resistant mechanisms available to the insect, the others being target-site insensitivity, reduced penetration through the cuticle and behavioural avoidance. By using a multi-variate (treatment, strain, time-points) experimental design, we hope to identify candidate resistance genes in *H. armigera* that warrant further examination and inform future studies.

4.2 Materials and Methods

This study was carried out in collaboration with members of the Land and Water Division at CSIRO, Black Mountain. Derivation and rearing of *H. armigera* strains (section 4.2.1) were carried out by Peter Hart and Claire Farnsworth. Peter Hart and I set up and performed the assay. I carried out the RNA extractions while Chris Coppin performed the preparation, quantification and normalisation of libraries for RNA-Seq (section 4.2.2 and Appendix A). Read assembly, alignment and preliminary analyses (section 4.2.3) were performed by Stephen Pearce. All other analyses in this thesis are my own work unless otherwise noted.

4.2.1 Samples and experimental design

The *H. armigera* strains used in this study were derived by Claire Farnsworth (PhD thesis, 2013). The strain originated from about 20 pairwise matings of individuals caught in the Namoi Valley during the 2008-2009 cropping season. About 80 F_2 individuals from these 20 matings were combined to form the BHA strain. BHA was selected for fenvalerate resistance through topical application of 1µl of a racemic mixture of fenvalerate solution in acetone on third-instar larvae. The concentration of fenvalerate applied was a multiple of the original discriminating dose (DD) which was 0.125µg/µl. Survival on 10×DD reached 100% by generation 2 while survival on 50×DD steadily climbed to 80% by generation 14 (Figure 4.1). Survival of BHA on 100×DD was variable as the health of the strain was compromised. A BHA-bc strain was derived by backcrossing BHA males to females from a laboratory general rearing strain when the BHA strain crashed due to inbreeding depression after 15 generations. BHA-bc was established from 160 backcross individuals (60 males and 100 females) that survived a 100×DD fenvalerate challenge and built up without selection for one generation. Two cohorts of BHA-bc were maintained in parallel: an unselected (UU) cohort and a selected (SS) cohort that was subjected to fenvalerate selection. Approximately 120 survivors of the 100×DD (12.5µg) were used to continue the strain after each round of fenvalerate selection in the SS cohort. At 100×DD, resistance in the SS cohort was below 20% at generations 15–16 then gradually climbed to over 80%

by generation 26. By contrast, survival of the UU cohort at generation 24 was only 36% at 50×DD (data not shown).

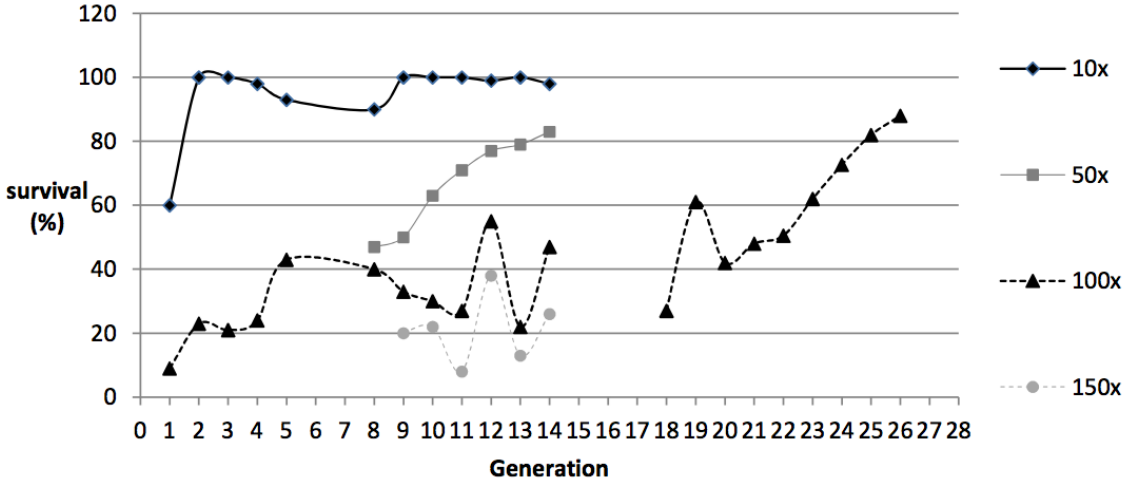


Figure 4.1: Percent survival of third instar BHA (generations 1–14) and BHA-bc-SS (generations 18–26) on various doses of fenvalerate over the course of fenvalerate selection. The original discriminating dose, 1×DD, was 0.125µg. (Figure courtesy of Claire Farnsworth, PhD thesis, 2013)

This study was performed approximately 15 generations after BHA-bc was established, whereby the *Cyp337b3* allele had reached fixation in the SS strain and approximately 70% frequency in the UU strain, with the remaining 30% of the UU strain carrying *Cyp337b1–Cyp337b2*. The genotypes were verified using the PCR screen outlined in the supplementary methods of Joußen *et al.* (2012). For this study, crosses were performed by mass-mating 22 males and 22 females within each strain, and the F_1 progeny were assayed. Progeny were collected as third-instar larvae and placed into trays containing artificial diet. Larvae from each cross were divided into two treatment groups:

(a) Unexposed (control): topical application of acetone only

(b) Exposed: topical application of 1.25µg of fenvalerate dissolved in acetone

A dose of 1.25µg (10×DD) was chosen based on pilot experiments that showed firstly, little or no mortality in both selected and unselected cohorts over a 24-hour period; and

secondly, noticeable developmental delays in only the unselected cohort beyond 24 hours. In other words, the aim was to choose a treatment that would elicit a differential response between the selected and unselected cohorts whilst maximising the number of survivors over a 24-hour period.

Larvae were collected and flash-frozen in liquid nitrogen at 4 timepoints: 1, 6, 12 and 24 hours post-treatment. The assay was set up and performed in triplicate (i.e. 3 biological replicates) on the same day, resulting in a total of 48 pools (3 $\times$ 2 strains $\times$ 2 treatments $\times$ 4 timepoints) where each pool contained a minimum of 10 larvae. Setup began at 8am, and collection times corresponded to approximately 10am, 3pm, 9pm and 9am the next day. A schematic of the experimental design is shown in Figure 4.2.

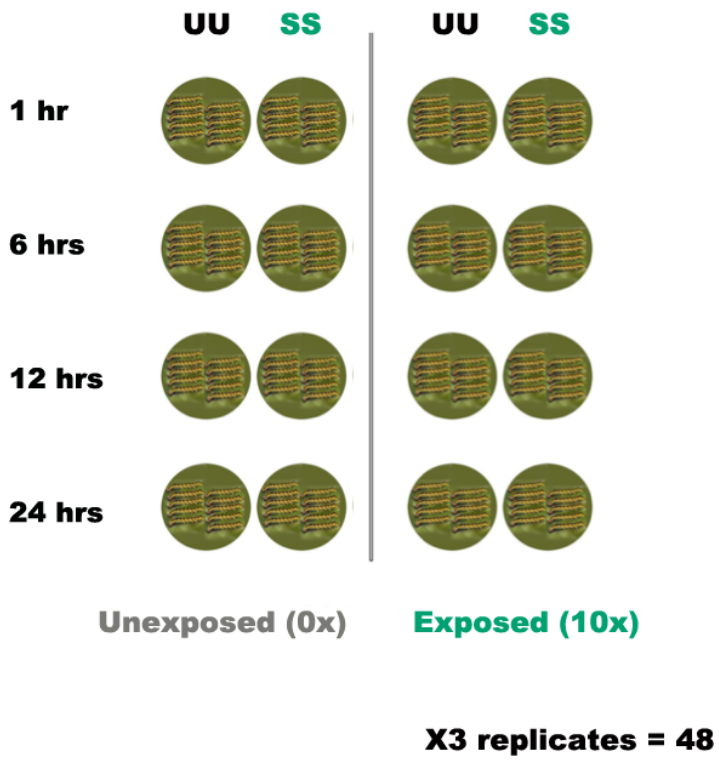


Figure 4.2: Schematic illustrating the samples and experimental design used in this study

4.2.2 Library preparation

A total of 48 libraries were generated, each with a unique barcode. Each library consisted of RNA extracted from 10 whole larvae. Total RNA was extracted using a guanidinium thiocyanate-phenol-chloroform protocol according to the manufacturer's instructions (TRIzol, Invitrogen). Libraries were prepared using the method of Langevin *et al.* (2013) with the protocol adapted for dual indexing and mRNA enrichment instead of rRNA depletion. Briefly, the method involves flanking the first-strand cDNA product with short tags which then serve as primer binding sites for incorporating full-length sequencing adapters and barcodes during second-strand synthesis. The first tag is appended onto the random hexamer that primes reverse transcription. The second tag, which has a different sequence, is attached to the 3' end of the completed first strand cDNA by taking advantage of the terminal nucleotidyl transferase and template switching activity of the reverse transcriptase. By attaching different tags to each end of the first strand cDNA, strand directionality information is retained. A second feature of this method is the use of a quantitative PCR assay to determine the number of cycles required for optimal amplification in the final library enrichment step, as over-cycling may introduce artefacts such as primer concatemers. Detailed protocols for mRNA isolation, first-strand cDNA synthesis and barcoding are included in Appendix A. Reaction yields were quantified using a KAPA Library Quantification Kit for Illumina Platforms according to the manufacturer's instructions.

4.2.3 Read assembly, alignment and preliminary analyses

Sequencing was performed on an Illumina NextSeq 500 platform using 75-cycle single-end reads across six runs. Reads were trimmed using `Trimmomatic` (Bolger *et al.*, 2014) to remove adapter sequences and low-quality reads. Reads that passed the filter were aligned to the *H. armigera* genome assembly (Pearce *et al.*, 2017) using the subread aligner implemented in the `R subread` package (Liao *et al.*, 2013), with a maximum of three mismatches allowed. The best scoring alignment for each read was reported. The read count table (read summarisation) was generated at the gene level by taking the number

of reads per library that overlapped with the predicted transcripts using `featureCounts` (Liao *et al.*, 2014).

4.2.4 Testing for differentially-expressed genes

Analyses of differential gene expression were performed using `edgeR` under the GLM (generalised linear model) framework (Robinson *et al.*, 2010). Two sets of filtering criteria were initially considered for the read count table prior to testing for differential gene expression across samples:

(1) `rowSums(cpm(y)>1)>=3` (low-stringency filter)

Only rows containing a cpm (counts per million) of greater than 1 in at least three samples were included. This filtering criterion is regularly employed in transcriptome studies whereby rows are included if they have sufficient read counts (using an arbitrary threshold for cpm, but typically 1) in at least n samples, where n is the number of biological replicates. The filter allows for zero values such as in cases where a gene is expressed in some treatments but not others, and it also increases the sensitivity towards detecting differential expression in genes that are lowly-expressed. However, this filter resulted in a BCV (biological coefficient of variation) value of 0.81. As the BCV represents the uncertainty with which the true abundance of a gene varies between biological replicates, a lower value is preferable. Increasing the cpm threshold only marginally reduced the BCV value, which provided the motivation for exploring a more stringent set of filtering criteria.

(2) `rowSums(cpm(y)>1)==96` (high-stringency filter)

Only rows containing a cpm of greater than 1 in all samples were included. This filter provided a BCV value of 0.38, which falls within the range of values typically observed in experimental designs that involve organisms from non-identical genetic backgrounds (McCarthy *et al.*, 2012). However, the stringency of this filter may have decreased its sensitivity towards detecting differentially-expressed genes. Data presented in the Results section is based on this filtering threshold as it was deemed preferable to take a more conservative approach.

Normalisation of library sizes between samples was performed using the recommended TMM (trimmed mean of M-values) method in the `calcNormFactors` function. The design matrix was constructed by defining a coefficient for the expression level of each group with no intercept i.e. `model.matrix(~0+group)` where `group` consisted of the 16 combinations of the strain-timepoint-treatment variables in this experiment.

To identify DE genes between unexposed and exposed cohorts within each strain at each timepoint (Table 4.3A in the Results section), the data was first divided into two subsets, with each subset corresponding to the libraries for each strain. Within each subset, a comparison between the unexposed and exposed cohort was performed for the relevant timepoint. Table 4.1 below illustrates the contrast parameters used to identify DE genes between unexposed and exposed cohorts in each strain at 1 hour; DE genes at the other three timepoints were identified in a similar manner.

Strain-Timepoint-Treatment combination	**Contrast sets**	
	UU	**SS**
SS-1-Exposed	0	1
SS-1-Unexposed	0	-1
SS-6-Exposed	0	0
SS-6-Unexposed	0	0
SS-12-Exposed	0	0
SS-12-Unexposed	0	0
SS-24-Exposed	0	0
SS-24-Unexposed	0	0
UU-1-Exposed	1	0
UU-1-Unexposed	-1	0
UU-6-Exposed	0	0
UU-6-Unexposed	0	0
UU-12-Exposed	0	0
UU-12-Unexposed	0	0
UU-24-Exposed	0	0
UU-24-Unexposed	0	0

Table 4.1: Contrast sets to illustrate identification of DE genes between exposed and unexposed cohorts in each strain at 1 hour post-exposure

To identify DE genes between strains at each timepoint under exposed and unexposed

conditions (Table 4.3B in the Results section), the UU strain was set as the reference and the SS strain was compared to the UU strain at each timepoint, with the analysis for exposed and unexposed cohorts performed separately. Table 4.2 below illustrates the contrast parameters used to identify DE genes between the two strains at 1 hour in the unexposed cohort; DE genes at the other three timepoints and for the exposed cohort were identified in a similar manner.

Strain-Timepoint-Treatment combination	**Contrast set for SS relative to UU**
SS-1-Exposed	0
SS-1-Unexposed	1
SS-6-Exposed	0
SS-6-Unexposed	0
SS-12-Exposed	0
SS-12-Unexposed	0
SS-24-Exposed	0
SS-24-Unexposed	0
UU-1-Exposed	0
UU-1-Unexposed	-1
UU-6-Exposed	0
UU-6-Unexposed	0
UU-12-Exposed	0
UU-12-Unexposed	0
UU-24-Exposed	0
UU-24-Unexposed	0

Table 4.2: Contrast set to illustrate identification of DE genes between strains at 1 hour in unexposed cohorts only

4.3 Results

All 48 libraries contained in excess of 10 million reads, and no systematic biases were observed in the distribution of library sizes (Figure 4.3). Of the 800 million reads obtained, approximately 21% of reads could not be mapped (not present in the reference or mapped to more than one location). A total of 17,089 transcripts which could be mapped to sequences with a functional annotation in the *H. armigera* genome (Pearce *et al.*, 2017) were obtained. All results presented in this section are based on the high-stringency filtering criteria (cpm>1 in all samples). Under these criteria, a total of 5663 out of the 17,089 transcripts (33%) were retained. Only transcripts showing differential expression with a false discovery rate (FDR) of below 0.05 are reported.

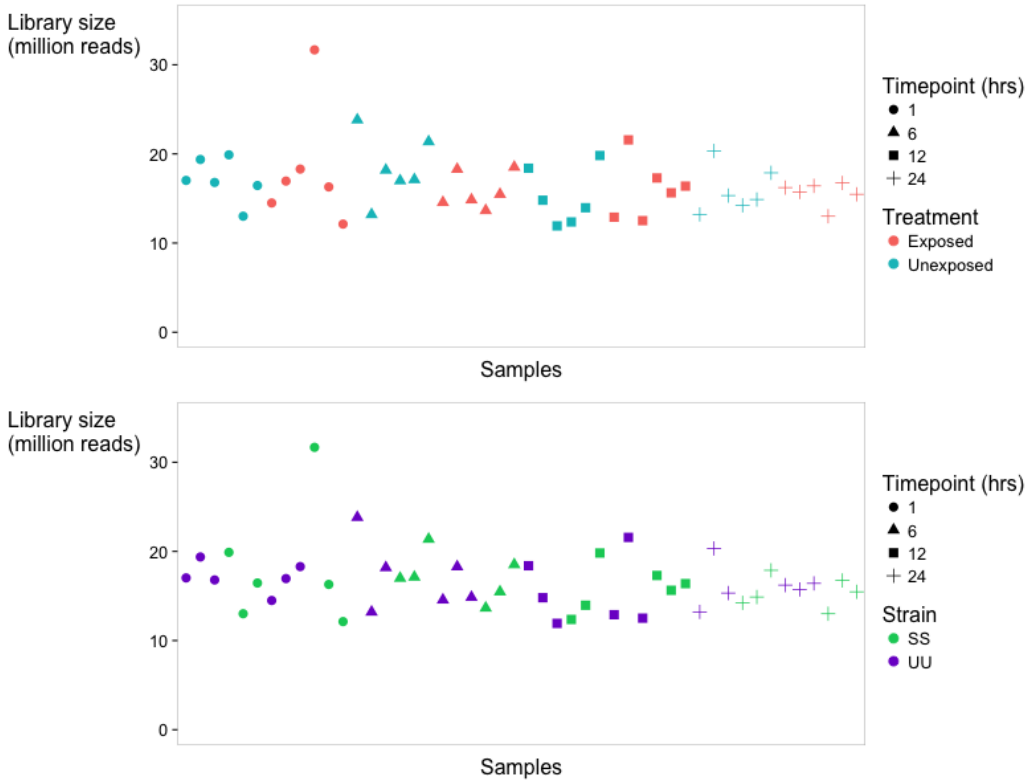


Figure 4.3: Distribution of library sizes (n=48) by timepoint, treatment (above) and strain (below)

4.3.1 Response at different timepoints

Overview

The three-way dimensionality of the experimental design allows us to analyse differential gene expression between (i) different timepoints, (ii) exposed and unexposed cohorts, which includes the response to induction; and (iii) selected and unselected strains, which includes the response to selection. An overview of the data reveals that both the selected and unselected strains respond to induction by fenvalerate, although in varying degrees. In the first analysis (Table 4.3A), unexposed and exposed cohorts are compared within a strain, and each strain is analysed separately to assess its response to induction. In both strains, there are more genes that are upregulated than downregulated in the exposed cohort. At the 1-hour timepoint, no differentially-expressed (DE) genes can be detected in either strain. Six hours post-exposure, a total of 1785 genes are found to be differentially expressed in the unselected (UU) strain compared to only 125 DE genes in the selected (SS) strain. There is a decrease in the number of genes responding to induction at the 12-hour timepoint in both strains but this dip appears to be temporary. At the 24-hour timepoint, both strains show an increase in the number of DE genes compared to 12 hours before. In the selected strain, the highest number of DE genes is observed at the 24-hour timepoint, unlike the unselected strain which peaks at the 6-hour timepoint.

When the analysis is performed with respect to the differences between strains which includes the selection response (Table 4.3B), a different pattern is observed. Using the unselected strain as the reference, the number of DE genes in the selected strain was quantified in a pairwise manner at each timepoint, with exposed and unexposed cohorts analysed separately. In the unexposed cohort of the selected strain, a moderate number of DE genes can be detected at the 1-hour timepoint. This result is not inconsistent with the response observed in the previous analysis (Table 4.3A) as it reflects constitutive differences between strains i.e. differential expression is detectible even at this early timepoint because basal transcription levels of these genes in the selected strain differ from the unselected strain. The number of DE genes decreases at the 6-hour timepoint, followed by an increase at the 12-hour timepoint. This is followed by a second dip at the 24-hour

A

Strain	No. of DE genes between Exposed and Unexposed cohorts at 1 hr	6 hrs	12 hrs	24 hrs
UU-up	0	916	76	285
UU-down	0	869	26	178
Total	0	1785	102	463
SS-up	0	66	41	94
SS-down	0	59	34	76
Total	0	125	75	170

B

Strain/ Treatment	No. of DE genes in SS strain relative to UU at 1 hr	6 hrs	12 hrs	24 hrs
Unexposed-up	32	4	135	58
Unexposed-down	24	7	145	27
Total	56	11	280	85
Exposed-up	4	27	31	19
Exposed-down	1	4	41	20
Total	5	31	72	39

Table 4.3: Number of genes that are differentially expressed
A) between exposed and unexposed cohorts within each strain
B) between strains under exposed and unexposed conditions
at 1 hour, 6 hours, 12 hours and 24 hours post-treatment.

For (A), the directionality refers to upregulation or downregulation in the exposed cohort relative to the unexposed. For (B), it refers to upregulation or downregulation in the SS strain relative to the UU strain.

timepoint. In the exposed cohort, only five DE genes are reported in the selected strain at the 1-hour timepoint. The number of DE genes then increases gradually over six hours, reaching a peak of 72 genes at the 12-hour timepoint. Similarly to the unexposed cohort, a dip occurs at the 24-hour timepoint with fewer DE genes reported at this stage.

To what extent are DE genes shared between timepoints and between strains?

The next level of analysis pertains to identifying a robust set of genes that warrant further study. We chose to focus on genes that appeared in more than one subset of the data – given that the dataset could be dissected in multiple ways, the presence of a gene in each subset could, to some extent, be considered an independent occurrence. For instance, if a gene was found to be differentially expressed at three different timepoints, this would be deemed a more robust candidate than one which was only found to be DE at one timepoint. The analysis was thus guided by this definition of 'robust', with a particular focus on genes that fell in the intersections between timepoints and between strains. In the first dataset where unexposed and exposed cohorts are compared within a strain, two questions that can be asked are: Which genes show consistent and robust patterns of DE at sufficiently high levels that they can be detected across all timepoints; and to what extent do the same genes appear in both the selected and unselected strains? A series of Venn diagrams (Figure 4.4) shows that there are 20 genes which continue to be differentially expressed in the UU strain at 6, 12 and 24 hours post-treatment, whereas the SS strain carries three such genes. No genes were found to be differentially expressed at the 1-hour timepoint in either strain. The list of 20 DE genes in the UU strain is primarily (75%) composed of enzyme-coding genes and includes genes encoding for two members of the detoxification gene families (UGT40M1 and CYP4G26), two serine protease-like proteins and a juvenile hormone epoxide hydrolase (Table 4.4). One of these 20 genes is downregulated (CG6459, also known as P32 in *D. melanogaster*). Of the three DE genes that occur across all three timepoints in the SS strain, one is downregulated. A gene which encodes for farnesyl pyrophosphate synthase (HaOG206975) is present in both the UU and SS strains.

The Venn diagrams in Figure 4.4 also show that 100, 21 and 89 DE genes are shared across both the UU and SS strains at the 6, 12 and 24-hour timepoints respectively. (The full list of genes is available in Appendix B.1, ranked by expression level of the SS strain; the selection of genes presented in this section pertain to some interesting features of the dataset discussed below.) At the 6-hour timepoint, 47% of the 100 DE genes here are up-

regulated. Amongst the upregulated genes are genes encoding for regulation of circadian rhythm (*timeless* and *cryptochrome*), an esterase, CCE033 and three CYPs: CYP9A17, CYP341B7 and CYP333B3 (Table 4.5). The downregulated genes are enriched for genes encoding 'housekeeping' proteins such as ribosome biogenesis proteins, ribosomal proteins, elongation factors and translation initiation factors. The list of DE genes at 12 hours has genes encoding for enzymes involved in detoxification (CYP4G8), tyrosine catabolism (4-hydroxyphenylpyruvate dioxygenase, HPPD) and lipid metabolism (Lipase61, fatty acid synthase-like protein). Only four of the 21 genes here are downregulated, and three of these four are housekeeping genes. At 24 hours, approximately 30% of the 89 DE genes here are downregulated. There are nine genes encoding for members of the detoxification gene families (CYP4AU7, CYP4G9, CYP4G26, CYP6AE19, CYP333A1, ABCG1, GSTϵ16, UGT33J1, UGT40M1), all of which are upregulated. A suite of genes related to development and lipid metabolism are also present in this list including genes involved in the regulation of juvenile hormone and genes encoding for cuticular proteins of which two are downregulated.

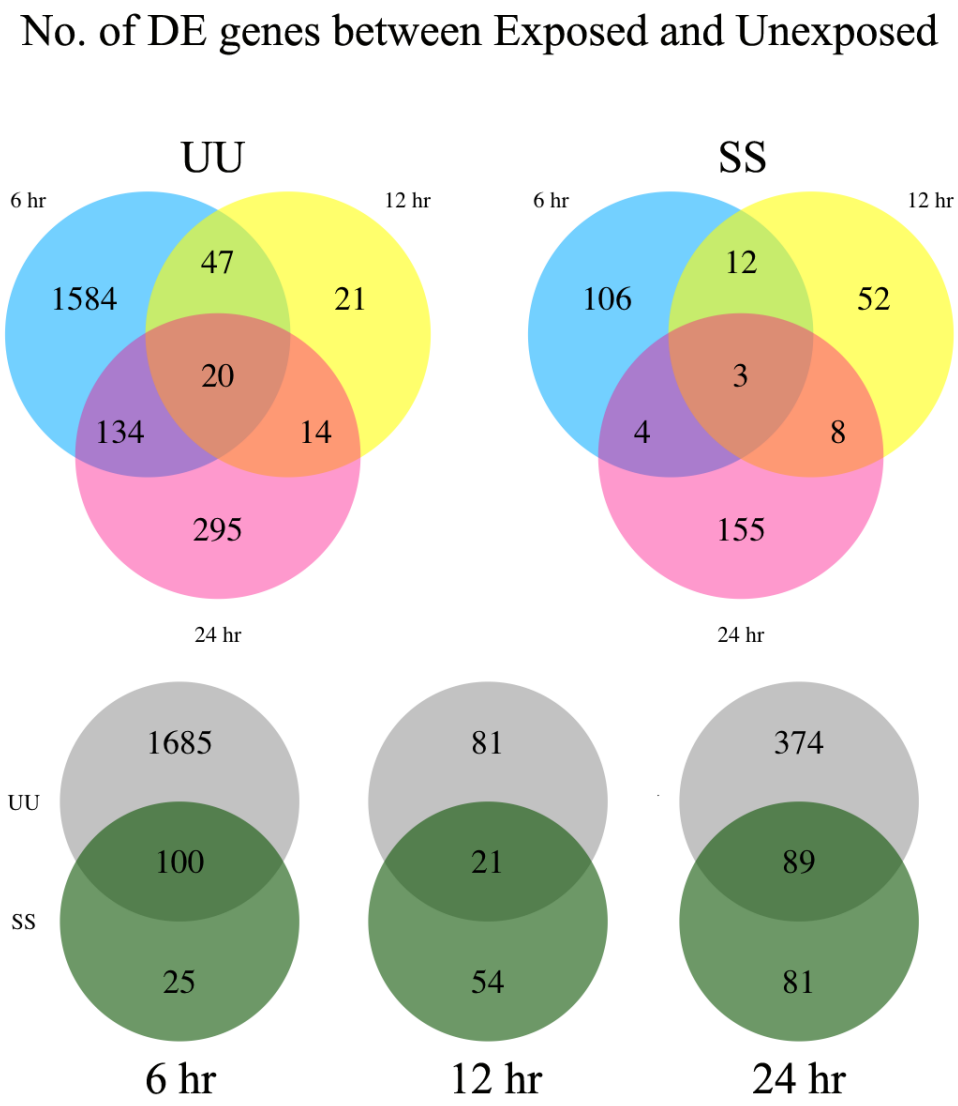


Figure 4.4: Venn diagrams illustrating the number of DE genes that are shared between timepoints and between strains in the comparison of exposed and unexposed cohorts within a strain (Table 4.3A). The top row depicts the number of DE genes shared across timepoints for the UU (left) and SS (right) strains while the bottom row depicts the number of DE genes that are in common between the UU and SS strains at each timepoint

	Gene annotation	HaOG	Log fold-change at 6 hrs	12 hrs	24 hrs
	UU strain				
1	BMORI:fumarylacetoacetase	HaOG215692	1.87	2.69	1.50
2	HarUGT40M1 ALT:HarUGT-14	HaOG200264	1.74	2.21	4.22
3	BMORI:fatty acid synthase-like	HaOG207602	1.72	2.42	2.83
4	CYP4G26-Ha	HaOG200072	1.62	2.08	2.08
5	BMORI:uncharacterized protein LOC101745021	HaOG206713	1.62	1.85	3.12
6	BMORI:trans-1,2-dihydrobenzene-1,2-diol dehydrogenase-like isoform X2	HaOG211544	1.48	2.61	2.52
7	DMELA:Q7KAK2 Farnesyl pyrophosphate synthase (Dimethylallyltransferase)	HaOG206975	1.42	1.23	1.35
8	BMORI:retinal dehydrogenase 1-like	HaOG216699	1.15	1.92	1.72
9	DMELA:Q9VSY0 Cuticular protein 67B	HaOG206673	1.14	1.64	1.80
10	BMORI:cuticular protein RR-1 motif 11 precursor	HaOG202657	1.13	1.27	1.59
11	serine protease easter-like	HaOG207183	1.13	1.68	1.41
12	DMELA:P52034 6-phosphofructokinase BMORI:6-phosphofructokinase-like isoform X2	HaOG216825	1.09	1.33	1.25
13	serine protease easter-like	HaOG208325	1.08	1.61	1.13
14	DMELA:Q7K537 GH14316p BMORI:selenium-binding protein 1-like isoform X1	HaOG205510	1.07	1.40	2.54
15	DMELA:Q95U46 GH07925p BMORI:acyl-coenzyme A dehydrogenase	HaOG201695	0.99	1.72	1.39
16	DMELA:P40320 S-adenosylmethionine synthase BMORI:S-adenosylmethionine synthetase	HaOG206272	0.95	1.39	1.29
17	DMELA:Q9VB96 CG31075 aldehyde dehydrogenase (NAD) activity	HaOG207201	0.93	1.74	1.60
18	DMELA:A1ZBF3 Juvenile hormone epoxide hydrolase 2, isoform C	HaOG214591	0.90	1.60	1.12
19	DMELA:A1Z992 CG33138 hydrolase and 1,4-alpha-glucan branching enzyme activity	HaOG204678	0.88	1.46	1.25
20	DMELA:Q7JXC4 CG6459 P32	HaOG215212	-2.23	-1.18	-1.16
	SS strain				
1	DMELA:Q9XZ56 4E-binding protein THOR	HaOG209464	1.60	1.84	1.44
2	DMELA:Q7KAK2 Farnesyl pyrophosphate synthase (Dimethylallyltransferase)	HaOG206975	1.50	1.53	1.06
3	DMELA:Q9VU36 LP04985p mitochondrial ribosomal protein L20	HaOG206747	-1.41	-1.50	-1.34

Table 4.4: Genes that are differentially expressed at 6, 12 and 24-hour timepoints in the UU (unselected) and SS (selected) strains, shown as $\log_2$ fold-change values. Values refer to differential expression between exposed and unexposed cohorts within each strain. Negative values represent downregulation in the exposed cohort relative to the unexposed cohort.

Rank	Gene annotation	HaOG	LogFC in UU	LogFC in SS
	At 6 hours			
	Detoxification:			
1	CYP9A17-Ha	HaOG200110	4.23	3.44
5	CYP341B7-Ha	HaOG200058	1.55	2.66
13	CYP333B3-Ha	HaOG200024	1.61	1.99
20	HaCCE033	HaOG200147	1.46	1.88
	Circadian rhythm:			
36	DMELA:P49021 Protein timeless BMORI:timeless	HaOG217123	1.77	1.37
47	HSAPI:Q16526 Cryptochrome-1 BMORI:cryptochrome 2	HaOG201315	1.49	0.93
	Housekeeping:			
48	DMELA:Q9VM33 Elongation factor G, mitochondrial BMORI:elongation factor G, mitochondrial-like	HaOG208027	-1.39	-0.91
53	DMELA:P56538 Eukaryotic translation initiation factor 6 BMORI:eukaryotic translation initiation factor 6	HaOG209996	-1.27	-1.01
56	DMELA:Q7K0Y1 Ribosome biogenesis protein BOP1 homolog BMORI:ribosome biogenesis protein BOP1 homolog	HaOG211250	-1.37	-1.03
60	DMELA:A1Z9E3 Elongation factor Tu BMORI:elongation factor Tu	HaOG212171	-1.83	-1.07
69	DMELA:Q9VKQ3 Ribosome biogenesis protein WDR12 homolog BMORI:ribosome biogenesis protein WDR12 homolog	HaOG216297	-2.23	-1.14
74	DMELA:Q9VJ38 39S ribosomal protein L13, mitochondrial BMORI:39S ribosomal protein L13, mitochondrial-like	HaOG208604	-1.43	-1.20
76	DMELA:Q9VXB5 39S ribosomal protein L22, mitochondrial BMORI:39S ribosomal protein L22, mitochondrial-like	HaOG211549	-2.12	-1.21
79	DMELA:Q9VLK2 Ribosomal L1 domain-containing protein CG13096 BMORI:ribosomal L1 domain-containing protein CG13096-like	HaOG209092	-2.21	-1.22
80	DMELA:P41375 Eukaryotic translation initiation factor 2 subunit 2 BMORI:eukaryotic translation initiation factor 2 subunit 2	HaOG215665	-2.05	-1.23
88	DMELA:Q9GQ89 Eukaryotic initiation factor eIF2B alpha subunit BMORI:eIF2B-alpha protein	HaOG203919	-1.52	-1.45
94	DMELA:Q9U3Z7 NHP2-like protein 1 homolog BMORI:ribosomal protein L7Ae	HaOG217260	-2.52	-1.76
99	DMELA:Q8T9B2 SD09147p BMORI:mitochondrial ribosomal protein L54	HaOG209676	-1.66	-1.97
	At 12 hours			
2	HaLipase61	HaOG200588	2.10	2.12
6	CYP4G8-Ha	HaOG200074	1.88	1.76
8	BMORI:fatty acid synthase-like	HaOG207602	2.42	1.70
9	DMELA:Q961W1 4-hydroxyphenylpyruvate dioxygenase BMORI:4-hydroxyphenylpyruvate dioxygenase-like	HaOG211200	2.55	1.66
	Housekeeping:			
18	HSAPI:P54136 Arginine–tRNA ligase, cytoplasmic BMORI:arginine–tRNA ligase, cytoplasmic-like	HaOG202143	-1.07	-0.96
19	DMELA:Q9VJ38 39S ribosomal protein L13, mitochondrial BMORI:39S ribosomal protein L13, mitochondrial-like	HaOG208604	-1.32	-1.16

21	DMELA:Q9VXB5 39S ribosomal protein L22, mitochondrial BMORI:39S ribosomal protein L22, mitochondrial-like	HaOG211549	-1.28	-1.45
	At 24 hours			
	Detoxification:			
7	HarUGT33J1 ALT:HarUGT-23	HaOG200284	4.22	3.34
12	CYP4AU7-Ha-26	HaOG200069	2.97	3.06
13	CYP4G9-Ha	HaOG200075	1.92	3.00
17	HarUGT40M1 ALT:HarUGT-14	HaOG200264	4.22	2.75
25	CYP6AE19-Ha	HaOG200094	1.86	2.14
29	HaABCG1 ALT:HaABC-G-05-1-H	HaOG200346	2.42	2.05
30	CYP4G26-Ha	HaOG200072	2.08	2.03
37	HaGSTe16	HaOG200231	2.37	1.82
53	CYP333A1-Ha	HaOG200023	1.86	1.42
	Development:			
8	BMORI:juvenile hormone binding protein an-0921 precursor	HaOG204351	4.40	3.22
47	BMORI:juvenile hormone binding protein an-0128 precursor	HaOG204844	1.68	1.60
58	DMELA:A1ZBF3 Juvenile hormone epoxide hydrolase 2, isoform C	HaOG214591	1.12	1.15
1	BMORI:cuticular protein RR-1 motif 46 precursor	HaOG201905	3.49	3.98
2	BMORI:cuticular protein RR-1 motif 46 precursor	HaOG201907	4.57	3.79
3	BMORI:cuticular protein RR-1 motif 46 precursor	HaOG201908	3.93	3.72
27	BMORI:cuticular protein glycine-rich 10 precursor	HaOG215934	1.61	2.08
34	BMORI:cuticular protein RR-1 motif 42 precursor	HaOG201900	1.59	1.87
35	BMORI:cuticular protein RR-1 motif 11 precursor	HaOG202657	1.59	1.83
36	DMELA:Q9VSY0 Cuticular protein 67B	HaOG206673	1.80	1.82
87	DMELA:Q8I0P8 Cuticular protein 65Av	HaOG214287	-1.73	-3.14
89	BMORI:cuticular protein RR-1 motif 32 precursor	HaOG201888	-3.41	-4.10
	Lipid metabolism:			
6	BMORI:fatty acid synthase-like	HaOG207602	2.83	3.43
20	HaLipase61	HaOG200588	2.75	2.53
50	BMORI:putative fatty acyl-CoA reductase CG5065-like	HaOG210892	1.32	1.47
54	BMORI:elongation of very long chain fatty acids protein AAEL008004-like	HaOG206661	2.60	1.37
55	BMORI:elongation of very long chain fatty acids protein AAEL008004-like	HaOG206662	1.83	1.33
65	DMELA:E1JHE4 Fatty acid (Long chain) transport protein, isoform C BMORI:fatty acid transport protein	HaOG210925	-0.98	-1.23

Table 4.5: DE genes that are present in both UU and SS strains – only a subset of genes are discussed and listed here from the set of 100, 21 and 89 DE genes that are present in both strains at the 6, 12 and 24-hour timepoints respectively. Rank number refers to the order in which each gene appears in the full list (Appendix B.1) ranked by level of differential expression in the SS strain. The level of differential expression in each strain is shown as $\log_2$ fold-change values.

An analysis of the overlaps between genes can also be performed with respect to the strain comparison presented in the overview (Table 4.3B). In this analysis, the number of DE genes in the selected strain was quantified at each timepoint using the unselected strain as the reference. Similarly, the question of which genes show consistent and robust patterns of DE across all timepoints can be asked here. For the purpose of identifying genes that show a constitutive difference between the selected and unselected strains, only the unexposed cohort is analysed here as the effects of induction cannot be disentangled from strain differences in the exposed cohort. A Venn diagram (Figure 4.5) shows that only one gene is differentially expressed across all timepoints in both strains, a gene which codes for a microvitellogenin-like protein (HaOG216755). A total of 38 genes are differentially expressed at two or more timepoints, nine of which are downregulated in the SS strain (Table 4.6). A gene encoding for a Twist-related protein is upregulated at the 1-hour timepoint but downregulated at the 12-hour timepoint. Three members of the detoxification gene families are amongst the 38 genes in this list: UGT40M1, GSTϵ16 and CYP6AN3.

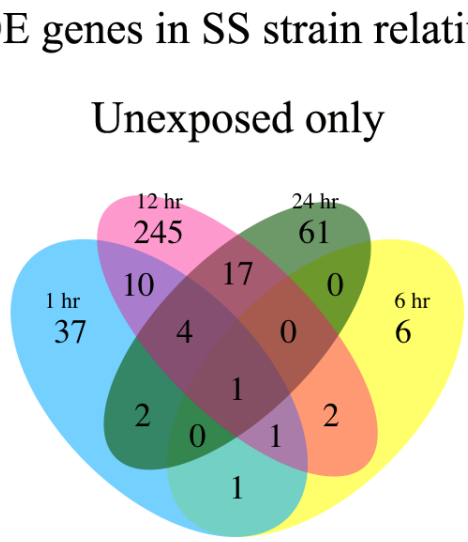


Figure 4.5: Venn diagram illustrating the number of DE genes in the SS strain relative to the UU strain that are shared between timepoints (Table 4.3B) for the unexposed cohort

	Gene annotation	HaOG	LogFC at			
			1 hr	6 hrs	12 hrs	24 hrs
1	BMORI:neurogenic locus notch homolog protein 1-like	HaOG212206	3.36		2.91	3.07
2	BMORI:microvitellogenin-like	HaOG216755	2.96	2.19	2.97	3.12
3	DMELA:Q7KTA2 CG33306	HaOG201413	2.13		2.10	
4	DMELA:A1Z6X6 CG1707	HaOG209351	1.79		2.19	1.65
5	DMELA:Q9VH09 CG3999	HaOG206725	1.75		1.51	
6	HSAPI:Q8WVJ9 Twist-related protein 2 BMORI:twist protein	HaOG217056	1.72		-1.59	
7	BMORI:probable ATP-dependent RNA helicase ddx17-like	HaOG212161	1.69		1.25	
8	DMELA:Q962N6 Flavin-containing monooxygenase FMO-1 BMORI:flavin-dependent monooxygenase FMO2 precursor	HaOG207733	1.68		1.47	
9	DMELA:Q8SXK3 RE16411p BMORI:2-amino-3-ketobutyrate coenzyme A ligase, mitochondrial-like	HaOG206931	1.51	1.91		
10	BMORI:neuroglian-like	HaOG210592	1.50			1.54
11	HSAPI:Q96G03 Phosphoglucomutase-2 BMORI:phosphoglucomutase-2-like	HaOG206851	1.39		1.74	1.46
12	DMELA:Q9W1I7 CG5554	HaOG208672	1.35		1.37	
13	HSAPI:P54105 Methylosome subunit pICln BMORI:methylosome subunit pICln-like	HaOG209379	1.25		1.73	
14	HarUGT40M1 ALT:HarUGT-14	HaOG200264		2.02	2.89	
15	DMELA:Q8IMH5 Niemann-Pick type C-2h, isoform B BMORI:promoting protein precursor	HaOG204615			2.50	1.75
16	HaGSTe16	HaOG200231			2.45	1.74
17	BMORI:hemolymph juvenile hormone binding protein precursor	HaOG204364			2.18	2.04
18	CYP6AN3-Ha	HaOG200099			2.11	2.80
19	DMELA:Q95U32 CG31344	HaOG207554		1.64	1.92	
20	HSAPI:Q9UHK6 Alpha-methylacyl-CoA racemase BMORI:alpha-methylacyl-CoA racemase-like isoform X1	HaOG203554			1.81	1.58
21	DMELA:Q9VGJ9 Heme oxygenase	HaOG213764			1.79	1.69
22	HSAPI:P26885 Peptidyl-prolyl cis-trans isomerase FKBP2 BMORI:peptidyl-prolyl cis-trans isomerase FKBP2-like	HaOG202690			1.73	1.82
23	DMELA:Q9VMC6 CG9547	HaOG216270			1.71	2.33
24	DMELA:Q9VQS4 Spindly BMORI:protein Spindly-like	HaOG206676			1.54	1.70
25	DMELA:Q9W3N9 CG10932	HaOG205801			1.50	1.15
26	BMORI:monocarboxylate transporter 14-like, partial	HaOG206815			1.44	1.79
27	BMORI:elongation of very long chain fatty acids protein AAEL008004-like	HaOG206662			1.29	1.50
28	BMORI:fatty acid-binding protein, heart-like	HaOG210112			1.10	1.31

29	DMELA:Q24492 Replication protein A 70 kDa DNA-binding subunit BMORI:replication protein A1	HaOG216188			1.10	1.06
30	BMORI:uncharacterized protein LOC101740603	HaOG204753			-1.14	-1.15
31	DMELA:Q9VKZ1 Protoheme IX farnesyltransferase, mitochondrial BMORI:protoheme IX farnesyltransferase, mitochondrial-like	HaOG206426			-1.33	-1.67
32	DMELA:P20028 DNA-directed RNA polymerase I subunit RPA2 BMORI:DNA-directed RNA polymerase I subunit RPA2-like	HaOG203697			-1.37	-1.37
33	DMELA:Q9VCD0 Glutamyl-tRNA(Gln) amidotransferase subunit B, mitochondrial BMORI:glutamyl-tRNA amidotransferase subunit B	HaOG202767	-1.16		-1.37	
34	BMORI:U3 small nucleolar RNA-associated protein 14 homolog A-like	HaOG202528	-1.20		-1.29	
35	no annotation	HaOG214947	-1.66		-1.90	
36	DMELA:Q9V3U1 CG3655, isoform A	HaOG206136	-1.71		-1.64	-1.71
37	DMELA:Q8IND1 Arpc3A, isoform D	HaOG207575	-1.92	-1.67	-1.35	
38	BMORI:uncharacterized protein PFB0145c-like	HaOG215026	-2.64			-2.95

Table 4.6: Subset of DE genes in the selected strain relative to the unselected strain, unexposed cohort – only genes that are differentially expressed at two or more timepoints are listed here. Genes are listed in order of level of differential expression at the 1-hour timepoint. The level of differential expression at each timepoint is shown as $\log_2$ fold-change values.

4.3.2 Response to induction and selection

Following on from the previous analyses, a union of the sets of DE genes at different timepoints shows that a total of 2203 genes in the UU and SS strains combined respond to the treatment, of which 252 genes are shared between both strains (Table 4.7A). Comparing between strains, a total of 387 genes are found to be differentially expressed in the unexposed SS strain relative to the UU strain. Table 4.7B draws on these sums to broadly categorise all the genes in this dataset into four categories through a contingency table: genes that are differentially expressed between treatments and strains (TS); between treatments but not strains (T); between strains but not treatments (S); and genes that show no differential responses between treatments nor strains (N). The 2203 genes that show a differential response between treatments could largely be considered inducible, but of the 387 genes that show a differential response between strains, only a subset may be responding to selection while the rest could be due to drift in the small population sizes kept in the laboratory. The genes that show a differential response between both treatments and strains are therefore of particular interest as they are more likely to have been selected for. Of the 219 genes that fall into this category, there are 18 members of the detoxification gene families (discussed further below). Other notable genes include genes encoding for haemolymph juvenile hormone binding protein precursors, mucin-like proteins, cuticular proteins and genes involved in lipid metabolism. Table 4.8 illustrates some of the genes in this dataset while Appendix C.1 lists all 219 genes in this category.

A

No. of DE genes	Strain UU	Strain SS	UU and SS (union)	UU and SS (intersection)
Between treatments (exposed and unexposed cohorts)	2115	340	2203	252
Between strains (unexposed SS strain relative to UU)	–	387	–	–

B

No. of genes that are	**DE between treatments**	**non-DE between treatments**	
DE between strains	219	168	387
non-DE between strains	1984	3292	5276
	2203	3460	**5663**

Table 4.7: A) Number of genes that are differentially expressed between treatments and between strains. Only the unexposed cohort was used to identify DE genes between strains.
B) Contingency table showing the four categories of DE genes. The total number of transcripts (5663) corresponds to the number of transcripts that passed the filtering criteria.

	HaOG	Gene annotation
		Detoxification
71	HaOG200023	CYP333A1-Ha
72	HaOG200072	CYP4G26-Ha
73	HaOG200094	CYP6AE19-Ha
74	HaOG200098	CYP6AN1-Ha
75	HaOG200099	CYP6AN3-Ha
186	HaOG200346	HaABCG1 ALT:HaABC-G-05-1-H
187	HaOG200341	HaABCH2 ALT:HaABC-H-26-2-H
188	HaOG200131	HaCCE017
189	HaOG200140	HaCCE026
190	HaOG200189	HaCCE107
191	HaOG200215	HaGSTe02
192	HaOG200231	HaGSTe16
193	HaOG200235	HaGSTo02
194	HaOG200239	HaGSTs01
195	HaOG200248	HaGSTs09
196	HaOG200249	HaGSTs10
199	HaOG200267	HarUGT33T1 ALT:HarUGT-20
200	HaOG200264	HarUGT40M1 ALT:HarUGT-14
		Cuticular proteins
12	HaOG215601	BMORI:cuticular protein glycine-rich 13 precursor
13	HaOG210770	BMORI:cuticular protein glycine-rich 6 precursor
14	HaOG201905	BMORI:cuticular protein RR-1 motif 46 precursor
		Juvenile hormone binding protein precursors
22	HaOG204363	BMORI:hemolymph juvenile hormone binding protein precursor
23	HaOG204364	BMORI:hemolymph juvenile hormone binding protein precursor
		Mucin-like proteins
28	HaOG215196	BMORI:mucin-17-like
29	HaOG212338	BMORI:mucin-2-like
30	HaOG214737	BMORI:mucin-5AC-like isoform X5
		Lipid metabolism
8	HaOG207595	BMORI:apolipoprotein D-like
17	HaOG206659	BMORI:elongation of very long chain fatty acids protein 2-like
18	HaOG206662	BMORI:elongation of very long chain fatty acids protein AAEL008004-like
19	HaOG210112	BMORI:fatty acid-binding protein, heart-like
148	HaOG211110	DMELA:Q9V496 Apolipophorins BMORI:apolipophorins isoform X2

Table 4.8: A selection of genes that are differentially expressed between treatments and between strains. Numbers in the leftmost column refer to the order in which they appear in the full list of 219 genes (Appendix C.1)

4.3.3 Contribution of the detoxification gene families

An aim of this study is to assess if members of the detoxification gene families are enriched in the set of DE genes. The five gene families combined – 114 CYPs, 97 CCEs, 42 GSTs, 46 UGTs and 54 ABCs – comprise 2% of the 17,000 protein-coding genes annotated in the *H. armigera* genome (Pearce *et al.*, 2017) and an *a priori* expectation is that the proportion of detoxification genes in the list of DE genes will exceed 2%. A count of the number of detoxification gene family members in each of the four classes of genes – differentially expressed between treatments and strains (TS); between treatments but not strains (T); between strains but not treatments (S); and genes that show no differential responses between treatments nor strains (N) – provides some evidence to support this hypothesis (Table 4.9). Of the 5663 transcripts that passed the filtering criteria, 1.7% consists of genes from the detoxification families, which approximates their proportion in the *H. armigera* genome. Of these 95 detoxification genes that passed the filtering criteria, the ABCs are particularly enriched, with 24 of their 54 members included here. Overall, the detoxification genes comprise over 8% of the TS class of genes. The GSTs are enriched in the TS class, with 6 members contributing to the 18 detoxification genes here, whereas most of the CYPs and UGTs are found in the T class. The detoxification families are also enriched (4.1%) in the S class of genes, but only account for 1.8% of the T class. One possibility is that inducible genes are also more likely to respond to selection, and their presence is compensated for by their over-representation in the TS class of genes.

	TS	**T**	**S**	**N**	**Total**
No. of genes in this class	219	1984	168	3292	5663
CYPs	5	11	1	5	22 (114)
CCEs	3	6	3	10	22 (97)
GSTs	6	3	1	7	17 (42)
ABCs	2	11	0	11	24 (54)
UGTs	2	5	2	1	10 (46)
Sum of detox. family genes	18	36	7	34	95
As % of genes in this class	8.2	1.8	4.1	1.0	1.7

Table 4.9: Proportion of CYPs, CCEs, GSTs, ABCs and UGTs that fall into the following four classes:
genes that are differentially expressed between treatments and strains (TS);
between treatments but not strains (T);
between strains but not treatments (S); and
genes that are not differentially expressed between treatments nor strains (N).
Figures in brackets in the rightmost column refer to the number of members for each gene family in the *H. armigera* genome

4.4 Discussion

The three-way dimensionality of the experimental design allowed the data to be dissected in several ways, providing a rich dataset with the potential to provide insights on several biological questions. An overview of the DE genes between exposed and unexposed cohorts over the time course (Table 4.3A) reveals that no differential expression is detectible at the 1-hour timepoint. This suggests that at this early timepoint, there has been insufficient time to allow for a differential response to the treatment to be detected. A second observation is that the increase in transcriptional activity between 1 and 6 hours appears to be followed by a dip in the induction response around the 12-hour timepoint. This profile is reminiscent of that observed by Willoughby *et al.* (2006) who compared the induction responses of *D. melanogaster* to eight xenobiotic compounds. Using real-time quantitative PCR, they quantified changes in mRNA levels at several timepoints over a 24-hour period and found that an initial induction peak occurred around 4 hours post-exposure; this was followed by a dip and an eventual increase to levels that exceeded the 4-hour peak as they approached the 24-hour mark. They hypothesised that the genes responding to the treatment could be metabolising or sequestering the xenobiotic in the first few hours post-exposure, thus lowering levels of the xenobiotic in the organism after the initial increase in transcriptional activity. A similar phenomenon could be responsible for the patterns observed in our data whereby the 6-hour timepoint could be (or is close to being) the initial induction peak.

A third observation is that the unselected (UU) strain shows a larger induction response – the UU strain has higher numbers of DE genes than the SS strain at each timepoint. This supports the idea that the unselected strain is exhibiting acute stress and increasing transcriptional activity of an arsenal of genes in response to the early stages of insecticide exposure. One intriguing possibility is that a 'master regulator' gene is induced by the stress of xenobiotic exposure in the UU strain but not in the SS. Misra *et al.* (2011) observed that up to 70% of the genes induced by xenobiotic compounds in *D. melanogaster* responded to expression of *CncC* (cap 'n' collar isoform C), a *Drosophila* orthologue of *Nrf2* which is a transcription factor from the leucine zipper family. Using transgenic

lines, they found that overexpressing *CncC* was sufficient to induce expression of six detoxification genes (*Cyp6a2, Cyp6a8, Cyp6a21, GSTδ2, GSTδ7* and *Jheh1*, juvenile hormone epoxide hydrolase 1) in the absence of xenobiotics. A similar mechanism could be present in *H. armigera* where the transcriptional activity of several genes in response to xenobiotics and/or oxidative stress can be traced to a single trans-acting factor, with a loss of function preventing the normal transcriptional response to xenobiotics from taking place in the SS strain.

In the overview of the DE genes between strains (Table 4.3B), the expression profile of the unexposed cohort provides evidence of constitutive differences between the selected and unselected strains – at the 1-hour mark, 56 DE genes are detectible, in contrast to the complete absence of DE genes at this timepoint in the previous analysis. The presence of DE genes at this early stage is consistent with differences in the basal transcription levels of these genes between the SS and UU strains. A question remains as to whether or not the DE genes reported in this analysis are relevant to fenvalerate resistance – it is not immediately apparent why the numbers of DE genes should vary over the time course in the absence of the insecticide. One possibility is that a subset of these constitutive differences are unrelated to the resistance phenotype, and that some genes simply have differential activity that align with the circadian rhythm of the organism. Somewhat counter-intuitively, the exposed cohort has fewer DE genes than the unexposed cohort over the time course, with the exception of the 6-hour timepoint. A possible explanation is that the response to induction in both strains has reduced the net effect of the differences between strains, making it more difficult to detect differential expression under our stringent filtering criteria.

In our dataset of DE genes between the exposed and unexposed cohorts (Table 4.4), genes encoding for serine protease-like proteins were found to be upregulated upon exposure to fenvalerate. Proteolytic enzymes have been reported to be upregulated in studies of differential gene expression between DDT and/or permethrin-resistant and susceptible strains of *A. gambiae*, *D. melanogaster* and houseflies (Ahmed *et al.*, 1998; Pedra *et al.*, 2004; Vontas *et al.*, 2005). The authors of these studies speculate on several possibilities

such as the involvement of protein catabolism in the induction of detoxification enzymes, putative roles of proteolytic enzymes in regulation through effecting changes in protein conformations and protein turnover, or protein catabolism as a means of regulating energy utilisation under stress. A gene encoding for juvenile hormone epoxide hydrolase (*Jheh*) was also upregulated in this dataset. *Jheh* has been found to be induced by multiple xenobiotics in *Trichoplusia ni*, including compounds without epoxides (Anspaugh and Roe, 2005).

Some interesting patterns can also be seen in the lists of genes that are shared between the UU and SS strains (Table 4.5) – at the 6-hour timepoint, a suite of housekeeping genes are downregulated in both strains. Inhibiting the translation of non-essential housekeeping proteins could be a stress response that allows the cell to conserve energy and divert resources towards adaptively responding to stressors (Roelofs *et al.*, 2010; Aufauvre *et al.*, 2014). *Cyp9a17* was also amongst the list of upregulated DE genes at this timepoint, which is consistent with a previous study that found it was upregulated in response to the pyrethroid insecticide deltamethrin, along with gossypol and phenobarbital (Zhou *et al.*, 2010a). Two genes involved in regulation of circadian rhythm, *crytochrome* and *timeless* were also upregulated at this timepoint, and it is possible that they could affect the activity of xenobiotic-metabolising enzymes at different times of day (Hooven *et al.*, 2009; Beaver *et al.*, 2010). Future studies would ideally control for this variation, perhaps by replicating the experiment at different times of the day.

At the 12 and 24-hour timepoints, genes involved in tyrosine catabolism, lipid metabolism and development are upregulated. 4-hydroxyphenylpyruvate dioxygenase (HPPD) is a key enzyme in tyrosine catabolism that metabolises 4-hydroxyphenylpyruvate into homogentisate. In silkworm larvae, carbamate insecticides have been shown to increase the concentration of hydroxyphenyl compounds in the haemolymph, so disrupting the function of HPPD could increase the amount of toxic metabolites circulating in the insect haemolymph (Sugiyama *et al.*, 1981). Another contribution of tyrosine to the resistance phenotype could be in reducing penetration of insecticides through the cuticle, as tyrosine serves as a precursor for compounds involved in cuticle sclerotization (Kramer and

Hopkins, 1987). The upregulation of genes involved in lipid metabolism and cuticular proteins is supported by studies in *Anopheles* which implicate these pathways in cuticular hydrocarbon synthesis (Jones *et al.*, 2013; Toé *et al.*, 2015). One explanation revolves around the possibility that the upregulation of these genes is a response to topical application of the insecticide as the insect attempts to reduce penetration of xenobiotics through the cuticle. Genes encoding enzymes involved in lipid metabolic pathways could also be effective targets for insect control because of their roles in maintaining the fat body, which is an organ that is unique to insects and has many metabolic functions including xenobiotic metabolism and energy storage (Arrese and Soulages, 2010). A second explanation, unrelated to insecticide exposure, is that the organism is undergoing developmental changes as it prepares to enter the next larval stage. Expression levels of *Jheh* vary over the course of larval development, as do cuticular proteins (Anspaugh and Roe, 2005; Togawa *et al.*, 2008; Liang *et al.*, 2010).

Comparing the unexposed SS strain to the UU strain (Table 4.6) reveals one gene that is differentially expressed across all four timepoints, a gene encoding for a microvitellogenin-like protein. Vitellogenin is a female-specific protein that is a major constituent of egg proteins. In most insects, vitellogenin is synthesised by the fat body and transported to the ovaries through the haemolymph. Microvitellogenin is a small protein that shares many similarities with vitellogenin (Kawooya *et al.*, 1986). A gene encoding for a haemolymph juvenile hormone binding protein (hJHBP) precursor, HaOG204364 is also upregulated in this dataset. HaOG204364 is also present in the list of genes that respond to induction, along with another hJHBP precursor, HaOG204363. Interestingly, both vitellogenin and hJHBP have been implicated in stress responses. Bees with higher concentrations of vitellogenin in their haemolymph were found to be more resistant to oxidative stress and had longer lifespans (Seehuus *et al.*, 2006) while stressed *Manduca sexta* showed decreased levels of hJHBP and a corresponding increase in the availability of unbound juvenile hormone to induce developmental delays in response to stressors (Tauchman *et al.*, 2007). Our data could therefore be interpreted as detecting reduced levels of hJHBP in the UU strain relative to the SS strain rather than an upregulation of hJHBP in the SS strain, which is consistent with the idea that the UU strain is exhibiting a larger stress response

while the SS strain is relatively unstressed. However, a limitation of the study design was the lack of replicate populations which would provide evidence for the contribution of a gene to the resistance phenotype. Differences between the selected and unselected strains could be attributable to drift rather than positive selection, so future study designs would ideally have replicate populations subjected to a selection regime without overly compromising population sizes, particularly in an organism like *H. armigera* which suffers from inbreeding depression. Our study design could also be further improved by distinguishing between different stages of the third-instar larvae i.e. whether the larvae were at an early, middle or late phase of third instar.

The classification of genes into the four categories (TS, T, S and N) reveals that larger numbers of DE genes are observed between treatments than between strains. In light of the limitations and caveats involved in the study design and data analyses, perhaps the most robust candidates that warrant further study are the class of genes that are differentially expressed both between treatments and strains (TS). A qualitative distinction between constitutive differences due to the selection response and those due to drift could be made by identifying DE genes that are present in both the between-treatments dataset (inducible genes) and the between-strains dataset – if a DE gene in the between-strains dataset is found to be DE in the between-treatments dataset, it appears more likely to have been selected for rather than simply being a consequence of drift. Using the example of the genes encoding for microvitellogenin and hJHBP precursor, the data could perhaps be interpreted as providing stronger support for the role of the latter in the resistance phenotype, whereas differential expression of the former could be attributed to drift. Similarly, the presence of genes involved in the lipid metabolism pathway and genes encoding for cuticular protein precursors in both the between-treatments and between-strains dataset suggests that these genes could be part of the selection response (in addition to being inducible) and contribute to the resistance phenotype.

Finally, our findings support the hypothesis that gene families involved in xenobiotic metabolism and detoxification such as the CYPs, CCEs, GSTs, UGTs and ABCs play a prominent role in the mechanisms underlying fenvalerate resistance. *Cyp337* was not

found to be amongst the list of genes in the filtered dataset. Nine members of the detoxification gene families are present in the set of 89 DE genes that are induced in both the UU and SS strains at the 24-hour timepoint. Across the whole dataset, these five gene families comprise approximately 2% of the 17,000 protein-coding genes in the *H. armigera* genome yet they are over-represented in the class of 'responsive' genes, particularly genes that are both inducible and selectable. Although they comprise only 1.8% of the inducible (T) class of genes, it is compensated for by their over-representation in the class of TS genes (8.2%). There are five CYPs in the TS class of genes, but it is not known if any of them have the ability to metabolise fenvalerate. As these gene families are found across a diverse range of organisms, further studies into the function of the individual members would provide valuable insight into the mechanisms underlying resistance across different orders of insect pests. An understanding of how different members of a metabolic pathway interact to confer resistance is also required to elucidate their roles in the induction and selection responses. Outside of the detoxification gene families, further investigation of juvenile hormone-binding proteins, cuticular proteins, tyrosine metabolism, lipid metabolic pathways and the effects of circadian rhythm on xenobiotic metabolism is also required to advance our understanding of their roles in insecticide resistance.

Chapter 5

General discussion

The aims of this thesis revolve around the premise that *H. armigera* is a target pest of insecticides in the field and is expected to be subjected to selective pressures that can leave behind signatures of selection in its genome. Candidate gene-finding for a trait of interest – in this case, insecticide resistance – can therefore be approached in several ways, of which two have been explored in this thesis: genomic scans for selective sweeps and comparative transcriptomics of populations that have been selected for and/or exposed to the insecticide. Chapter 2 explored patterns of variation using Australian populations to establish a baseline for future genome-wide scans of selective sweeps in *H. armigera*, while Chapter 3 assessed whether or not the findings of the previous chapter could be extended to non-Australian populations of *H. armigera*. Chapter 3 also assessed the extent of population differentiation between inter-continental samples and identified useful markers for characterising population structure. The genomic approach explored in Chapters 2 and 3 was complemented by a transcriptome study in Chapter 4 to identify candidate resistance genes by comparing gene expression between an unselected strain and a strain from the same genetic background that had undergone selection for fenvalerate resistance, and between exposed and unexposed cohorts. Here, I discuss some of the findings of this thesis, their implications, and future directions.

5.1 High nucleotide diversity and limited linkage disequilibrium in *Helicoverpa armigera* facilitates the detection of a selective sweep: Implications and future directions

In quantifying population genetic parameters such as nucleotide diversity, insertion-deletion frequency and the decay of LD, this study found that the *H. armigera* genome harboured high levels of diversity and high levels of recombination. It is not well understood how or why levels of genetic diversity vary between species, although Romiguier *et al.* (2014) report that life-history traits such as propagule size, fecundity and longevity appear to have some predictive value through their effects on effective population size, N_e. The range of genetic diversity levels can be relatively narrow despite large differences

in census population sizes across species (Leffler *et al.*, 2012). An upper or lower limit may exist to constrain the amount of variation that can be carried in a genome. The cost of carrying too many deleterious mutations in a population (genetic load) is detrimental, as is an inability to respond to change due to insufficient genetic variation (Crow, 1970). Intriguingly, high levels of diversity are observed in *H. armigera*, yet inbreeding depression, which is a consequence of homozygosity, is a recurrent problem amongst colonies of *H. armigera* maintained in the laboratory – it is not clear how the two phenomena can be reconciled. Studies of genetic load in *H. armigera* would provide insights into these questions as well as the factors underlying its reproductive success and fitness in the field.

The rapid decay of LD in *H. armigera* highlights some design considerations for genome-wide studies in this species. Designs that genotype only a subset of the variants using arrays of 'tag' SNPs (or 'SNP chips') such as those typically used in human studies are likely to be unpropitious, and genome imputation approaches will be limited. The genome size of *H. armigera* (and insect genomes in general) makes whole-genome sequencing feasible, but genome assembly will be challenging due to the high frequencies of single nucleotide as well as insertion-deletion polymorphisms. However, the low levels of LD suggest that identification of causal variants in GWAS and scans for selective sweeps could be easier because each association will be confined to a narrower 'window' (genomic region) containing only a small number of candidate genes.

A surprising outcome of this study was the identification of a locus that exhibited signs of a selective sweep. The study was conceived to survey patterns of variation at neutrally-evolving loci, hence the focus on intronic sequences and the EPIC marker design; it was, in effect, an attempt to answer the question 'What does neutral look like in *H. armigera*?'. However, the data revealed that the *Cyp303a1* locus was characterised by two divergent haplogroups, *Ins200* and *Del200*, with the *Del200* haplogroup displaying the classic signs of recent positive selection. Sequencing of the coding regions did not reveal any non-synonymous sites distinguishing between the two alleles, and it is not known what the function of *Cyp303a1* is in *H. armigera*. In *D. melanogaster*, the gene has been found to be essential for mechano- and chemosensation and is expressed only in the sensory bristles

(Willingham and Keil, 2004). Loss-of-function mutants had a diminished proboscis extension reflex (PER), a response scored as present or absent when leg chemoreceptors were stimulated with sugars or other tastants. In *H. armigera*, CYP303A1 was found to be expressed in the antennae (Pearce *et al.*, 2017), which suggests that it is likely to have a chemosensory function as antennae are the primary organs of olfaction in Lepidoptera. No expression was detected in the tissues associated with detoxification such as midgut, fatbody and Malphigian tubules. A functional study of *Cyp303a1* and its alleles in *H. armigera* would shed light on its contribution to insecticide resistance, if any, and expand our understanding of CYPs and their diverse roles across insect taxa. Another avenue of interest could be to quantify the frequency of chromosomal aberrations in *H. armigera* and to investigate the likelihood that the lack of recombination between the *Ins200* and *Del200* alleles is the result of a chromosomal inversion.

5.2 Population differentiation between Australian and Chinese *Helicoverpa armigera* occurs in distinct blocks on the Z chromosome: Implications and future directions

The previous study using Australian *H. armigera* provided a picture of a high-diversity genome, in a highly vagile species. To assess if these findings could be generalised, it was imperative that non-Australian populations be characterised in a similar manner – as Australasia was postulated to be the centre of origin for the *Helicoverpa* lineage, it was possible that Australian populations of *H. armigera* were atypical in their genetic diversity. To that end, samples were collected from two sites in China, located 700km apart. Less than 1% of the variation in the samples discriminated between individuals from the two collection sites, which was consistent with the high dispersal ability of the species. The Chinese populations also harboured levels of diversity similar to that of the Australian populations, and a rapid decay of LD. High genetic variation thus appeared to be fairly typical of *H. armigera*. Another finding in this study was a consistent signal of a negative Tajima's *D*. As the signal was distributed across multiple loci, the most parsimonious interpretation is that there has been a population expansion in the evolutionary history

of *H. armigera*. Future population genomic studies of *H. armigera* could incorporate different demographic models and employ more sophisticated statistical frameworks to disentangle the effects of selection from drift and date the expansion event.

Comparison of the samples from the two different countries revealed that the Chinese individuals could be distinguished from the Australian individuals, but only subtly – 88% of the variation could be attributed to variation within samples from the same country. The exception was the *Cyp303a1* locus where different haplotypes were predominant in each country. The discovery that the *Cyp303a1* locus was highly differentiated between Australian and Chinese *H. armigera* motivated a search for other loci that could be used as markers of population structure. A chromosome-wide scan identified several regions ('blocks') of high differentiation between Australian and Chinese populations, and these blocks formed distinct peaks (of high F_{ST} values) across the chromosome. That the *Cyp303a1* locus exhibits both population structure and signatures of a selective sweep raises the question as to whether or not these regions of high differentiation will show a similar phenomenon – resequencing of these loci in populations from different continents could uncover more evidence of positive selection in the *H. armigera* genome. One of these highly-differentiated regions contains members of the ABCB subfamily of ABC transporters which have been implicated in xenobiotic detoxification, so evidence of a selective sweep at this locus would provide strong support for its role in insecticide resistance.

The question of whether or not the *H. armigera* genome contains DNA introgressed from another species remains open; more data would be required to test the hypothesis, including identification of the source species. The introgression hypothesis presents some interesting possibilities – it could explain why Australian *H. armigera* appear to be a separate subspecies (*H. armigera conferta*) that differs from 'rest-of-the-world' *H. armigera* (Anderson *et al.*, 2016). An introgression event between *H. armigera* and a closely-related *Helicoverpa* species that is endemic to Australia would produce haplotypes that are unique to Australia, such as the *Cyp303a1 Del200* haplogroup. The presence of these blocks of high F_{ST} favour a model in which introgressed regions are concentrated in small stretches of high differentiation, rather than diffused in low-to-moderate amounts throughout the

genome. In a model proposed by Currat and Excoffier (2011), introgression coupled with population expansion could produce sweep-like patterns as genes introgressing in the invading population spread through the population. This model provides a backdrop for an alternative hypothesis: that the sweep-like signal at the *Cyp303a1 Del200* haplogroup is a result of demographic processes rather than selection. It would be interesting to see if more *Cyp303a1*-like loci can be found in the *H. armigera* genome, and to quantify the frequency of these patterns to determine the more parsimonious explanation.

Our study also confirmed instances of synteny between *B. mori* and *H. armigera*, consistent with previous reports in lepidopterans (Jiggins *et al.*, 2005; Yasukochi *et al.*, 2006; d'Alencon *et al.*, 2010). The study was conceived while assembly of the *H. armigera* genome was still underway, so coding sequences of the *B. mori* Z chromosome were used to identify putative Z-linked contigs in *H. armigera* through orthology – marker design thus relied upon assumptions of synteny in lepidopterans, and data collection was undertaken on that premise. The markers were only much later confirmed to be on the *H. armigera* Z chromosome, after completion of the *H. armigera* genome assembly and annotation. This auxiliary result supports the utility of synteny-based approaches in studies involving non-model lepidopterans as it may speed up the process of identifying insecticide resistance genes in pest organisms (Baxter *et al.*, 2011).

A caveat in both our studies is that the results may be peculiar to the Z chromosome. Sex chromosomes may be subjected to faster rates of evolution compared to the autosomes because recessive alleles are exposed in the hemizygous sex (Vicoso and Charlesworth, 2006; Sackton *et al.*, 2014). The symmetry of autosomal loci may buffer against the effects of selection on deleterious recessive alleles and drift as a result of larger effective population size (4/3 times that of Z-linked loci). However, a strength of this study design is the ability to empirically observe the different haplotypes resulting from recombination without the need for imputation – sequencing the sex chromosome in the hemizygous sex is a cost-beneficial way of overcoming the problem of inferring gametic phase from diploid sequences. Advances in single-molecule sequencing technology along with expected decreases in cost will presumably lead to the obsolescence of this approach in future.

5.3 Transcriptome analyses of the induction and selection response to fenvalerate in *Helicoverpa armigera*: Implications and future directions

A component that is lacking in genomic scans of selective sweeps is the link between genotype and phenotype. Signals of selection inferred purely from patterns in the DNA cannot directly be ascribed to a selective agent; a functional assay or study is required at some level to provide a biological context for the relevance of candidate loci to a particular trait. A comparative analysis of the transcriptomes of resistant and susceptible strains provides some validation of the relevance of candidate loci to insecticide resistance. While evidence of differential gene expression in itself may not be sufficient, it does moves us one step closer towards linking genotype to phenotype.

This study approached the identification of candidate genes by looking for differential gene expression between treatments (exposure to the insecticide) and between strains. In the induction response, genes respond when the substrate is encountered, so a comparison between exposed and unexposed cohorts identifies genes involved in this response. In the selection response, strains are expected to show constitutive differences in the expression of some genes due to the selection regime; a comparison between strains under unexposed conditions identifies genes involved in this response without the confounding influence of the induction response. Genes were assigned to one of four categories: responding to both induction and selection (TS), responding to induction but not selection (T), responding to selection but not induction (S), and responding to neither induction nor selection (N). The majority of the genes fell into the N class, while genes from the T class comprised the largest proportion of the three responsive classes. Some interesting candidates were revealed in the TS class of genes, and these genes form a good starting point for future studies as they are found to be differentially expressed in multiple 'slices' of the dataset. The xenobiotic response of genes involved in lipid metabolism is supported by the role of the fatbody which is analogous to the mammalian liver and has many functions in-

cluding xenobiotic metabolism. It also serves as an endocrine organ and as the site of synthesis for haemolymph proteins. Lipophorin, a major lipid carrier circulating in the haemolymph, has the ability to bind a range of xenobiotics including permethrin, dieldrin, DDT and malathion (Haunerland and Bowers, 1986). Lipid metabolism also serves to regulate energy utilisation when the insect is faced with stressors, provides cholesterol for synthesising steroid hormones including ecdysone and delivers energy to flight muscles and oocytes (Canavoso *et al.*, 2001; Arrese and Soulages, 2010).

The upregulation of genes involved in tyrosine metabolism is conceivable given that carbamate insecticides increase the concentration of hydroxyphenyl compounds in the insect haemolymph (Sugiyama *et al.*, 1981); upregulating HPPD could serve to reduce the amount of these toxic metabolites in the haemolymph. Another potential role for tyrosine in insecticide resistance is through cuticle sclerotization (Kramer and Hopkins, 1987), where resistant cohorts could have evolved to reduce penetration of xenobiotics through additional hardening of the cuticle. Copley (2000) offer an interesting hypothesis of pre-existing promiscuous enzymes being recruited to perform novel functions – they speculate that a glutathione-dependent enzyme from the tyrosine catabolism pathway, maleylacetoacetate (MAA) isomerase, may have been recruited to detoxify the pesticide pentachlorophenol (PCP) in *Sphingomonas chlorophenolica* (a Gram-negative aerobic bacteria) due to similarities in the active sites between the MAA isomerase and a dehalogenase (both of which are ζ-class GSTs) involved in the degradation of PCP. While this scenario seems quite idiosyncratic, it does spark some interesting possibilities which could lead to other avenues of research.

A variable which is often not considered in insecticide dose-response assays is the time of day at which assays are carried out. Circadian clocks have been found to regulate not only everyday physiological functions such as sleep-wake cycles, cellular function and metabolism, but also the activity of xenobiotic-metabolising enzymes (Hooven *et al.*, 2009; Beaver *et al.*, 2010). Beaver *et al.* (2010) found that disruption of the transcription factor *Pdp1*, which is controlled by the circadian clock, led to differences in the expression of *Cyp6a2, Cyp6g1,* α*-Esterase7* and *DHR96*, a nuclear receptor which has been

found to regulate expression of a large number of genes involved in xenobiotic metabolism (King-Jones *et al.*, 2006) – thus leading to differences in pesticide susceptibility. It is speculated that this variation in the expression of xenobiotic-metabolising enzymes may have evolved to anticipate exposure to plant allelochemicals and other compounds during daily feeding rhythms. An avenue for further research would be to investigate if expression of *CncC*, another central regulator of xenobiotic-metabolising enzymes (Misra *et al.*, 2011) also exhibits a similar response to the circadian clock. Hooven *et al.* (2009) also found that the variation in susceptibility to insecticides at different times of day was not significant when flies were maintained under constant light, so future study designs could use this approach to control for such variation.

The upregulation of genes encoding for cuticular proteins and juvenile hormone-related functions presents some intriguing questions: do they contribute to the resistance phenotype, or is the differential expression detected in this study a result of developmental changes taking place over the course of the 24-hour assay? The answers may not be mutually exclusive – for instance, juvenile hormone degradation is affected by stress and insects can delay progression to the next developmental stage in response to stressors (Rauschenbach *et al.*, 1996; Gruntenko *et al.*, 2000; Tauchman *et al.*, 2007). Similarly, cuticular proteins form part of the normal developmental pathway in insects but could also serve to reduce penetration of xenobiotics through the cuticle. A further consideration is that genes encoding for cuticular proteins and JH-related functions may also be governed by circadian rhythms, and/or may be part of the suite of genes under the control of a central regulator such as *DHR96* or *CncC*. More research would be required to understand the contributions of these genes in insecticide resistance.

This study assessed the contribution of five detoxification gene families to fenvalerate resistance, and the data support the hypothesis that these gene families are enriched in the set of genes responding to induction and selection. CYP303A1 was not among the CYPs found to be differentially expressed in this study, and there are several plausible reasons for its absence in this dataset: firstly, expression of CYP303A1 in the *H. armigera* antennae would indicate that it is unlikely to have a role in xenobiotic detoxification.

Secondly, if the gene was only expressed at very low levels in *H. armigera* (as it is in *D. melanogaster*), it is unlikely to have passed the stringent filtering criteria used in this study. Further, the sweep-like patterns were only observed in the *Del200* haplogroup – it is not clear if this intronic insertion-deletion polymorphism was segregating amongst the individuals used in this study. Genomic sequences would be required to ascertain if there are any associations between the CYP303A1 indel polymorphism and resistance.

A criticism that could be levelled at this study is the lack of replication of both the selection and control lines. As only a finite number of individuals will be carried through each generation, allele frequencies in subsequent generations are susceptible to the effects of drift. The use of replicate populations helps to disentangle the effects of drift from selection, particularly if the same genes are found to be differentially expressed in multiple replicates. However, even study designs that have rigorously taken into consideration the effects of effective population size, laboratory adaptation and the replication of selection and control lines have found it difficult to identify consistent genetic changes across replicates. Griffin *et al.* (2017) found that there were large differences in the number of loci estimated as contributing to the selection response, with some replicate lines showing up to ten times the number of candidate SNPs as other replicate lines in some cases. The authors concluded that these differences were the result of hitchhiking effects and different selection responses in the replicate lines. Selection also appeared to promote divergence rather than convergence of the replicate lines, and it was unclear if additional precautions to control for hitchhiking effects would have borne a different outcome.

Another limitation of this study was the variation between biological replicates that led to greater uncertainty surrounding the true abundance of a gene. The use of high-stringency filtering criteria to estimate levels of DE with a higher confidence comes at the expense of discarding data – only 33% of transcripts passed the filtering threshold. Ideally, a more sophisticated statistical technique would be employed in future study designs to retain a larger proportion of the dataset. One avenue that could be explored to reduce the variation between replicates is to further categorise the larvae into sub-categories of early, middle or late-stage third instar as the differences in developmental timing could

be contributing to the 'noise' in the data.

A QTL (quantitative trait loci) mapping approach will provide further insights in linking genotype to phenotype. There are several options for obtaining a quantitative measure of the resistance phenotype including time to death, median lethal dose (LD_{50}), knockdown time, fecundity and motility response (Denecke *et al.*, 2015). In a cross between the selected (resistant) and unselected (susceptible) strain, the segregating progeny resulting from an F_2 cross (or F_1 backcrossed to a parental strain) will display different mean values of the trait. Genome sequencing of the segregating progeny along with a scan for selective sweeps is expected to provide a valuable dataset to propel the search for novel candidates. Given that a major resistance factor has already been identified in the form of *Cyp337b3*, the use of parental strains that have fixed for *Cyp337b3* will allow genes of smaller effect to be elucidated. The high levels of recombination in *H. armigera* populations may prove to be helpful at pinpointing the causal locus within the QTL region. Nevertheless, functional studies involving positional cloning and targeted gene replacement would still be required to establish causality.

A further endeavour would be to build a repository of sequenced genotypes against which the transcriptomes could be compared to identify correlations between variation in gene expression and genotype i.e. eQTLs (expression quantitative trait loci). Mapping crosses have already been undertaken, and analysis of QTLs and eQTLs is expected to take place in the near future (J. Oakeshott, personal communications). Additionally, changes in cis- and trans-regulatory elements could be investigated through the use of reciprocal crosses, with the question: How does prior selection by an insecticide affect subsequent transcriptional responses, without the supposition that such changes are adaptive responses to insecticides. ChIP (chromatin immunoprecipitation) sequencing could also be undertaken to identify the binding sites of transcription factors.

5.4 Concluding remarks

The body of research presented in this thesis provides insights into the genome architecture of *H. armigera*, a major pest of agriculture globally. In the context of applied research, it identifies several candidate genes for pyrethroid resistance that warrant further investigation. It also identifies several avenues for future inquiries in population genomics and evolutionary biology.

References

Ahmad M, Denholm I and Bromilow RH (2006) Delayed cuticular penetration and enhanced metabolism of deltamethrin in pyrethroid-resistant strains of *Helicoverpa armigera* from China and Pakistan. *Pest Management Science* **62**(9):805–810

Ahmad M, Gladwell R and McCaffery A (1989) Decreased nerve sensitivity is a mechanism of resistance in a pyrethroid resistant strain of Heliothis armigera from Thailand. *Pesticide biochemistry and physiology* **35**(2):165–171

Ahmed S, Wilkins RM and Mantle D (1998) Comparison of proteolytic enzyme activities in adults of insecticide resistant and susceptible strains of the housefly M. domestica L. *Insect Biochemistry and Molecular Biology* **28**(9):629–639

Ahn SJ, Vogel H and Heckel DG (2012) Comparative analysis of the UDP-glycosyltransferase multigene family in insects. *Insect Biochemistry and Molecular Biology* **42**(2):133–147

Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Valimaki N, Paulin L, Kvist J, Wahlberg N, Tanskanen J, Hornett EA, Ferguson LC, Luo S, Cao Z, de Jong MA, Duplouy A, Smolander OP, Vogel H, McCoy RC, Qian K, Chong WS, Zhang Q, Ahmad F, Haukka JK, Joshi A, Salojarvi J, Wheat CW, Grosse-Wilde E, Hughes D, Katainen R, Pitkanen E, Ylinen J, Waterhouse RM, Turunen M, Vaharautio A, Ojanen SP, Schulman AH, Taipale M, Lawson D, Ukkonen E, Makinen V, Goldsmith MR, Holm L, Auvinen P, Frilander MJ and Hanski I (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun* **5**

Anderson CJ, Tay WT, McGaughran A, Gordon K and Walsh TK (2016) Population structure and gene flow in the global pest, *textitHelicoverpa armigera. Molecular Ecology*

Anspaugh DD and Roe RM (2005) Regulation of JH epoxide hydrolase versus JH esterase activity in the cabbage looper, Trichoplusia ni, by juvenile hormone and xenobiotics. *Journal of Insect Physiology* **51**(5):523–535

Anthony N, Unruh T, Ganser D and ffrench Constant R (1998) Duplication of the Rdl GABA receptor subunit gene in an insecticide-resistant aphid, Myzus persicae. *Molecular & general genetics: MGG* **260**(2-3):165–175

Arnemann JA, James WJ, Walsh TK, Guedes JVC, Smagghe G, Castiglioni E and Tay WT (2016) Mitochondrial DNA COI characterization of *Helicoverpa armigera* (Lepidoptera: Noctuidae) from Paraguay and Uruguay. *Genetics and molecular research: GMR* **15**(2)

Arrese EL and Soulages JL (2010) Insect Fat Body: Energy, Metabolism, and Regulation. *Annual Review of Entomology* **55**(1):207–225

Aufauvre J, Misme-Aucouturier B, Viguès B, Texier C, Delbac F and Blot N (2014) Transcriptome Analyses of the Honeybee Response to Nosema ceranae and Insecticides. *PLOS ONE* **9**(3):e91686

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA and Johnson EA (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* **3**(10):e3376

Bassi A, Rison JL and Wiles JA (2009) Chlorantraniliprole (DPX-E2Y45, Rynaxypyr®, Coragen®), A new diamide insecticide for control of codling moth (Cydia pomonella), Colorado potato beetle (Leptinotarsa decemlineata) and European grapevine moth (Lobesia botrana). *Zbornik predavanj in referatov 9. Slovenskega Posvetovanja o Varstvu Rastlin, Nova Gorica, Slovenije, 4-5 marec 2009* pp. 39–45

Battlay P, Schmidt JM, Fournier-Level A and Robin C (2016) Genomic and Transcriptomic Associations Identify a New Insecticide Resistance Phenotype for the Selective Sweep at the Cyp6g1 Locus of *Drosophila melanogaster*. *G3 (Bethesda, Md.)* **6**(8):2573–2581

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD and Blaxter ML (2011) Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism. *PLOS ONE* **6**(4):e19315

Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A, Beltran M, Perez-Espona S, Chamberlain N, Ferguson L, Clark R, Davidson C, Glithero R, Mallet J, McMillan WO, Kronforst M, Joron M, ffrench Constant RH and Jiggins CD (2010) Genomic hotspots for adaptation: The population genetics of Mullerian mimicry in the *Heliconius* melpomene clade. *PLoS Genet* **6**(2):e1000794

Beaver LM, Hooven LA, Butcher SM, Krishnan N, Sherman KA, Chow ESY and Giebultowicz JM (2010) Circadian Clock Regulates Response to Pesticides in Drosophila via Conserved Pdp1 Pathway. *Toxicological Sciences* **115**(2):513–520

Behere G, Tay W, Russell D, Heckel D, Appleton B, Kranthi K and Batterham P (2007) Mitochondrial DNA analysis of field populations of *Helicoverpa* armigera (Lepidoptera: Noctuidae) and of its relationship to *H.* zea. *BMC Evolutionary Biology* **7**(1):117

Behere GT, Tay WT, Russell DA, Kranthi KR and Batterham P (2013) Population genetic structure of the cotton bollworm *textitHelicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) in India as inferred from EPIC-PCR DNA markers. *PLoS One* **8**(1)

Bloomquist JR (1993) Toxicology, mode of action and target site-mediated resistance to insecticides acting on chloride channels. *Comparative Biochemistry and Physiology. C, Comparative Pharmacology and Toxicology* **106**(2):301–314

Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**(15):2114–2120

Brand CL, Kingan SB, Wu L and Garrigan D (2013) A Selective Sweep across Species Boundaries in Drosophila. *Mol Biol Evol* **30**(9):2177–2186

Brisson J, Nuzhdin S and Stern D (2009) Similar patterns of linkage disequilibrium and nucleotide diversity in native and introduced populations of the pea aphid, *Acyrthosiphon* pisum. *BMC Genetics* **10**(1):22

Browning SR and Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**(5):1084–1097

Brun-Barale A, Héma O, Martin T, Suraporn S, Audant P, Sezutsu H and Feyereisen R (2010) Multiple P450 genes overexpressed in deltamethrin-resistant strains of *Helicoverpa armigera*. *Pest. Manag. Sci.* pp. n/a–n/a

Buss DS and Callaghan A (2008) Interaction of pesticides with p-glycoprotein and other ABC proteins: A survey of the possible importance to insecticide, herbicide and fungicide resistance. *Pesticide Biochemistry and Physiology* **90**(3):141–153

Canavoso LE, Jouni ZE, Karnas KJ, Pennington JE and Wells MA (2001) Fat metabolism in insects. *Annual Review of Nutrition* **21**:23–46

Carino FA, Koener JF, Plapp FW and Feyereisen R (1994) Constitutive overexpression of the cytochrome P450 gene CYP6A1 in a house fly strain with metabolic resistance to insecticides. *Insect Biochemistry and Molecular Biology* **24**(4):411–418

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**(3):195–205

Charlesworth B, Coyne JA and Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist* **130**(1):113

Cho S, Mitchell A, Mitter C, Regier J, Matthews M and Robertson R (2008) Molecular phylogenetics of heliothine moths (Lepidoptera: Noctuidae: Heliothinae), with comments on the evolution of host range and pest status. *Systematic Entomology* **33**(4):581–594

Collart MA and Panasenko OO (2012) The CCR4-Not complex. *Gene* **492**(1):42–53

Common I (1953) The Australian species of Heliothis (Lepidoptera: Noctuidae) and their pest status. *Australian Journal of Zoology* **1**(3):319

Copley SD (2000) Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends in Biochemical Sciences* **25**(6):261–265

Crow JF (1970) Genetic Loads and the Cost of Natural Selection. In Ki Kojima (editor) *Mathematical Topics in Population Genetics*, no. 1 in Biomathematics, pp. 128–177. Springer Berlin Heidelberg. ISBN 978-3-642-46246-7 978-3-642-46244-3

Crow JF and Kimura M (1970) An introduction to population genetics theory. pp. xiv+591 pp.

Currat M and Excoffier L (2011) Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proceedings of the National Academy of Sciences* **108**(37):15129–15134

Czepak C, Albernaz KC, Vivan LM, Guimarães HO and Carvalhais T (2013) First reported occurrence of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) in Brazil. *Pesquisa Agropecuária Tropical* **43**(1):110–113

Daborn P, Boundy S, Yen J, Pittendrigh B and ffrench Constant R (2001) DDT resistance in Drosophila correlates with Cyp6g1 over-expression and confers cross-resistance to the neonicotinoid imidacloprid. *Molecular Genetics and Genomics* **266**(4):556–563

d'Alencon E, Sezutsu H, Legeai F, Permal E, Bernard-Samain S, Gimenez S, Gagneur C, Cousserans F, Shimomura M, Brun-Barale A, Flutre T, Couloux A, East P, Gordon K, Mita K, Quesneville H, Fournier P and Feyereisen R (2010) Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements. *Proceedings of the National Academy of Sciences* **107**(17):7680–7685

Daly JC (1993) Ecology and genetics of insecticide resistance in *Helicoverpa armigera*: Interactions between selection and gene flow. *Genetica* **90**(2):217–226

Daly JC and Fisk JH (1998) Sex-linked inheritance of endosulphan resistance in *Helicoverpa armigera*. *Heredity* **81**(1):55–62

Daly JC and Gregg P (1985) Genetic variation in Heliothis in Australia: Species identification and gene flow in the two pest species *H.* armigera (Hübner) and *H.* punctigera Wallengren (Lepidoptera: Noctuidae). *Bulletin of Entomological Research* **75**(01):169–184

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R and 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**(15):2156–2158

Davies TGE, Field LM, Usherwood PNR and Williamson MS (2007) DDT, pyrethrins, pyrethroids and insect sodium channels. *IUBMB Life* **59**(3):151–162

Dawkar VV, Chikate YR, Lomate PR, Dholakia BB, Gupta VS and Giri AP (2013) Molecular Insights into Resistance Mechanisms of Lepidopteran Insect Pests against Toxicants. *Journal of Proteome Research* **12**(11):4727–4737

Denecke S, Nowell CJ, Fournier-Level A, Perry T and Batterham P (2015) The Wiggle Index: An Open Source Bioassay to Assess Sub-Lethal Insecticide Response in Drosophila melanogaster. *PLOS ONE* **10**(12):e0145051

Dermauw W and Van Leeuwen T (2014) The ABC gene family in arthropods: Comparative genomics and role in insecticide transport and resistance. *Insect Biochemistry and Molecular Biology* **45**(Supplement C):89–110

Devonshire AL and Moores GD (1982) A carboxylesterase with broad substrate specificity causes organophosphorus, carbamate and pyrethroid resistance in peach-potato aphids (Myzus persicae). *Pesticide Biochemistry and Physiology* **18**(2):235–246

Duret L (2008) Neutral theory: The null hypothesis of molecular evolution. *Nature Education* **1**(1):218

Durigan MR, Corrêa AS, Pereira RM, Leite NA, Amado D, de Sousa DR and Omoto C (2017) High frequency of CYP337B3 gene associated with control failures

of *Helicoverpa armigera* with pyrethroid insecticides in Brazil. *Pesticide Biochemistry and Physiology*

Earl DA and vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**(2):359–361

Edi CV, Djogbénou L, Jenkins AM, Regna K, Muskavitch MAT, Poupardin R, Jones CM, Essandoh J, Kétoh GK, Paine MJI, Koudou BG, Donnelly MJ, Ranson H and Weetman D (2014) CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito Anopheles gambiae. *PLoS Genet* **10**(3):e1004236

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES and Mitchell SE (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **6**(5):e19379

Endersby NM, Hoffmann AA, McKechnie SW and Weeks AR (2007) Is there genetic structure in populations of *Helicoverpa* armigera from Australia? *Entomologia Experimentalis et Applicata* **122**(3):253–263

Epis S, Porretta D, Mastrantonio V, Comandatore F, Sassera D, Rossi P, Cafarchia C, Otranto D, Favia G, Genchi C, Bandi C and Urbanelli S (2014) ABC transporters are involved in defense against permethrin insecticide in the malaria vector Anopheles stephensi. *Parasites & Vectors* **7**:349

Evanno G, Regnaut S and Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**(8):2611–2620

Ewing G and Hermisson J (2010) MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**(16):2064–2065

Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, Yamaoka Y, Mégraud F, Otto K, Re-

ICHARD U, KATZOWITSCH E, WANG X, ACHTMAN M and SUERBAUM S (2003) Traces of Human Migrations in Helicobacter pylori Populations. *Science* **299**(5612):1582–1585

FANG QQ, CHO S, REGIER JC, MITTER C, MATTHEWS M, POOLE RW, FRIEDLANDER TP and ZHAO S (1997) A New Nuclear Gene for Insect Phylogenetics: DOPA Decarboxylase is Informative of Relationships within Heliothinae (Lepidoptera: Noctuidae). *Syst Biol* **46**(2):269–283

FEYEREISEN R (1999) Insect P450 Enzymes. *Annu. Rev. Entomol.* **44**(1):507–533

FEYEREISEN R (2005) Insect cytochrome P450. *Comprehensive Molecular Insect Science* **4**:1–77

FFRENCH CONSTANT RH, PITTENDRIGH B, VAUGHAN A and ANTHONY N (1998) Why are there so few resistance–associated mutations in insecticide target genes? *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **353**(1376):1685–1693

FIELD LM, DEVONSHIRE AL, FFRENCH-CONSTANT RH and FORDE BG (1989) Changes in DNA methylation are associated with loss of insecticide resistance in the peach-potato aphid Myzus persicae (Sulz.). *FEBS Letters* **243**(2):323–327

FIELD LM, DEVONSHIRE AL and FORDE BG (1988) Molecular evidence that insecticide resistance in peach-potato aphids (Myzus persicae Sulz.) results from amplification of an esterase gene. *Biochemical Journal* **251**(1):309–312

FISK JH and DALY JC (1989) Electrophoresis of *Helicoverpa armigera* (Hübner) and H. Punctigera (Wallengren) (Lepidoptera: Noctuidae): Genotype expression in eggs and allozyme variations between life stages. *Aust J Entomol* **28**(3):191–192

FITT GP (1989) *The Ecology of Heliothis Species in Relation to Agroecosystems*

FITT GP (1994) Cotton Pest Management: Part 3. An Australian Perspective. *Annual Review of Entomology* **39**(1):543–562

FITT GP (2003) Deployment and impact of transgenic Bt cotton in Australia. In *The economic and environmental impacts of Agbiotech*, pp. 141–164. Springer

Forrester NW, Cahill M, Bird LJ, Layland JK and others (1993) Management of pyrethroid and endosulfan resistance in *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Australia. *Bulletin of Entomological Research: Supplement Series* (Supplement 1)

Gahan LJ, Gould F and Heckel DG (2001) Identification of a gene associated with Bt resistance in Heliothis virescens. *Science (New York, N.Y.)* **293**(5531):857–860

Gahan LJ, Pauchet Y, Vogel H and Heckel DG (2010) An ABC Transporter Mutation Is Correlated with Insect Resistance to Bacillus thuringiensis Cry1Ac Toxin. *PLoS Genet* **6**(12):e1001248

Garud NR, Messer PW, Buzbas EO and Petrov DA (2015) Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLoS Genet* **11**(2):e1005004

Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T and Martin JF (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **12**(1):245

Goldsmith MR and Marec F (2009) *Molecular Biology and Genetics of the Lepidoptera*. CRC Press. ISBN 978-1-4200-6020-1

Good RT, Gramzow L, Battlay P, Sztal T, Batterham P and Robin C (2014) The Molecular Evolution of Cytochrome P450 Genes within and between Drosophila Species. *Genome Biol Evol* **6**(5):1118–1134

Gordon KHJ, Tay WT, Collinge D, Williams A and Batterham P (2010) Genetics and molecular biology of the major crop pest genus *Helicoverpa*. CRC Press. ISBN 978-1-4200-6014-0

Goudet J (2005) hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**(1):184–186

Gouy M, Guindon S and Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**(2):221–224

Grant DF and Hammock BD (1992) Genetic and molecular evidence for a trans-acting regulatory locus controlling glutathione S-transferase-2 expression in *Aedes aegypti*. *Molecular & general genetics: MGG* **234**(2):169–176

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, De La Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichter EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D and Pääbo S (2010) A Draft Sequence of the Neandertal Genome. *Science* **328**(5979):710–722

Greenstone MH, Stuart MK and Haunerland NH (1991) Using Monoclonal Antibodies for Phylogenetic Analysis: An Example from the Heliothinae (Lepidoptera: Noctuidae). *Annals of the Entomological Society of America* **84**(5):457–464

Griffin PC, Hangartner SB, Fournier-Level A and Hoffmann AA (2017) Genomic Trajectories to Desiccation Resistance: Convergence and Divergence Among Replicate Selected Drosophila Lines. *Genetics* **205**(2):871–890

Grubor VD and Heckel DG (2007) Evaluation of the role of CYP6B cytochrome P450s in pyrethroid resistant Australian *Helicoverpa* armigera. *Insect Molecular Biology* **16**(1):15–23

Gruntenko N, G Wilson T, Monastirioti M and Rauschenbach I (2000) Stress-reactivity and juvenile hormone degradation in *Drosophila melanogaster* strains having stress-related mutation. *Insect biochemistry and molecular biology* **30**:775–83

Gunning R, Devonshire A and Moores G (1995) Metabolism of Esfenvalerate by Pyrethroid-Susceptible and -Resistant Australian *Helicoverpa armigera* (Lepidoptera: Noctuidae). *Pesticide Biochemistry and Physiology* **51**(3):205–213

Gunning RV and Easton CS (1989) Pyrethroid resistance in Heliothis armigera (Hübner) collected from unsprayed maize crops in New South Wales, 1983–1987. *Australian Journal of Entomology* **28**(1):57–61

Gunning RV, Easton CS, Balfe ME and Ferris IG (1991) Pyrethroid resistance mechanisms in Australian *Helicoverpa armigera*. *Pestic. Sci.* **33**(4):473–490

Gunning RV, Easton CS, Greenup LR and Edge VE (1984) Pyrethroid Resistance in Heliothis armiger (Hübner) (Lepidoptera: Noctuidae) in Australia. *Journal of Economic Entomology* **77**(5):1283–1287

Guo Y, Shen YH, Sun W, Kishino H, Xiang ZH and Zhang Z (2011) Nucleotide diversity and selection signature in the domesticated silkworm, *Bombyx* mori, and wild silkworm, *Bombyx* mandarina. *Journal of Insect Science* **11**(1):155

Hama H, Kono Y and Sato Y (1987) Decreased Sensitivity of Central Nerve to Fenvalerate in the Pyrethoroid-Resistant Diamondback Moth, Plutella xylostella LINNE (Lepidoptera : Yponomeutidae). *Applied Entomology and Zoology* **22**(2):176–180

Han Y, Yu W, Zhang W, Yang Y, Walsh T, Oakeshott JG and Wu Y (2015) Variation in P450-mediated fenvalerate resistance levels is not correlated with CYP337B3 genotype in Chinese populations of *Helicoverpa armigera*. *Pesticide Biochemistry and Physiology* **121**(Supplement C):129–135

Hardwick DF (1965) The corn earworm complex. *Memoirs of the Entomological Society of Canada* **97**(Supplement S40):5–247

Harris C, Rousset F, Morlais I, Fontenille D and Cohuet A (2010) Low linkage disequilibrium in wild *Anopheles* gambiae s.l. populations. *BMC Genetics* **11**(1):81

Haunerland NH and Bowers WS (1986) Binding of insecticides to lipophorin and arylphorin, two hemolymph proteins of Heliothis zea. *Archives of Insect Biochemistry and Physiology* **3**(1):87–96

Head DJ, McCaffery AR and Callaghan A (1998) Novel mutations in the *para*-homologous sodium channel gene associated with phenotypic expression of nerve in-

sensitivity resistance to pyrethroids in Heliothine lepidoptera. *Insect molecular biology* **7**(2):191–196

Heckel D (2009) Molecular Genetics of Insecticide Resistance in Lepidoptera. In *Molecular Biology and Genetics of the Lepidoptera*, Contemporary Topics in Entomology. CRC Press. ISBN 978-1-4200-6014-0

Heckel DG, Gahan LJ, Daly JC and Trowell S (1998) A Genomic Approach to Understanding Heliothis and Helicoverpa Resistance to Chemical and Biological Insecticides. *Phil. Trans. R. Soc. Lond. B* **353**(1376):1713–1722

Hill WG and Robertson A (1968) Linkage disequilibrium in finite populations. *Theoret. Appl. Genetics* **38**(6):226–231

Hill WG and Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology* **33**(1):54–78

Holsinger KE and Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics* **10**(9):639–650

Hooven LA, Sherman KA, Butcher S and Giebultowicz JM (2009) Does the Clock Make the Poison? Circadian Variation in Response to Pesticides. *PLOS ONE* **4**(7):e6469

Huang HS, Hu NT, Yao YE, Wu CY, Chiang SW and Sun CN (1998) Molecular cloning and heterologous expression of a glutathione S-transferase involved in insecticide resistance from the diamondback moth, Plutella xylostella. *Insect Biochemistry and Molecular Biology* **28**(9):651–658

Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. *Genetics Research* **50**(03):245–250

Hudson RR, Bailey K, Skarecky D, Kwiatowski J and Ayala FJ (1994) Evidence for positive selection in the superoxide dismutase *(Sod)* region of *Drosophila* melanogaster. *Genetics* **136**(4):1329–1340

Hudson RR and Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**(1):147–164

Hudson RR, Slatkin M and Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**(2):583–589

Ji YJ, WU YC and ZHANG DX (2005) Novel polymorphic microsatellite markers developed in the cotton bollworm *Helicoverpa* armigera (Lepidoptera: Noctuidae). *Insect Science* **12**(5):331–334

Jiggins CD, Mavarez J, Beltran M, McMillan WO, Johnston JS and Bermingham E (2005) A genetic linkage map of the mimetic butterfly *Heliconius* melpomene. *Genetics* **171**(2):557–570

Jombart T and Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**(21):3070–3071

Jones CM, Haji KA, Khatib BO, Bagi J, Mcha J, Devine GJ, Daley M, Kabula B, Ali AS, Majambere S and Ranson H (2013) The dynamics of pyrethroid resistance in Anopheles arabiensis from Zanzibar and an assessment of the underlying genetic basis. *Parasites & Vectors* **6**:343

Joußen N, Agnolet S, Lorenz S, Schöne SE, Ellinger R, Schneider B and Heckel DG (2012) Resistance of Australian *Helicoverpa armigera* to fenvalerate is due to the chimeric P450 enzyme CYP337B3. *Proceedings of the National Academy of Sciences* **109**(38):15206–15211

Joußen N and Heckel DG (2016) Resistance Mechanisms of *Helicoverpa armigera*. *SpringerLink* pp. 241–261

Kamvar ZN, Tabima JF and Grünwald NJ (2014) *Poppr* : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**:e281

Kawooya JK, Osir EO and Law JH (1986) Physical and chemical properties of microvitellogenin. A protein from the egg of the tobacco hornworm moth, Manduca sexta. *Journal of Biological Chemistry* **261**(23):10844–10849

Kelley JL and Swanson WJ (2008) Positive Selection in the Human Genome: From Genome Scans to Biological Significance. *Annu. Rev. Genom. Human Genet.* **9**(1):143–160

Kennaugh L, Pearce D, Daly J and Hobbs A (1993) A Piperonyl Butoxide Synergizable Resistance to Permethrin in *Helicoverpa armigera* Which Is Not Due to Increased Detoxification by Cytochrome P450. *Pesticide Biochemistry and Physiology* **45**(3):234–241

Kimura M (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press. ISBN 978-0-521-31793-1

King EG and Coleman RJ (1989) Potential for biological control of Heliothis species. *Annual Review of entomology* **34**(1):53–75

King-Jones K, Horner MA, Lam G and Thummel CS (2006) The DHR96 nuclear receptor regulates xenobiotic responses in Drosophila. *Cell Metabolism* **4**(1):37–48

Kramer KJ and Hopkins TL (1987) Tyrosine metabolism for insect cuticle tanning. *Archives of Insect Biochemistry and Physiology* **6**(4):279–301

Kriticos DJ, Ota N, Hutchison WD, Beddow J, Walsh T, Tay WT, Borchert DM, Paula-Moreas SV, Czepak C and Zalucki MP (2015) The Potential Distribution of Invading *Helicoverpa armigera* in North America: Is It Just a Matter of Time? *PLOS ONE* **10**(3):e0119618

Kuchler K (2011) The ABC of ABCs: multidrug resistance and genetic diseases. *FEBS J*

Labbé R, Caveney S and Donly C (2011) Genetic analysis of the xenobiotic resistance-associated ABC gene subfamilies of the Lepidoptera. *Insect Molecular Biology* **20**(2):243–256

Lakey A, Labeit S, Gautel M, Ferguson C, Barlow DP, Leonard K and Bullard B (1993) Kettin, a large modular protein in the Z-disc of insect muscles. *The EMBO journal* **12**(7):2863–2871

Langevin SA, Bent ZW, Solberg OD, Curtis DJ, Lane PD, Williams KP, Schoeniger JS, Sinha A, Lane TW and Branda SS (2013) Peregrine. *RNA Biology* **10**(4):502–515

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, Fang S, Nista PM, Holloway AK, Kern AD, Dewey CN, Song YS, Hahn MW and Begun DJ (2012) Genomic variation in natural populations of *textitDrosophila melanogaster*. Genetics* **192**(2):533–598

Lanning CL, Fine RL, Corcoran JJ, Ayad HM, Rose RL and Abou-Donia MB (1996) Tobacco budworm P-glycoprotein: biochemical characterization and its involvement in pesticide resistance. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1291**(2):155–162

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21):2947–2948

Laster ML and Hardee DD (1995) Intermating compatibility between North American *Helicoverpa* zea and *Heliothis* armigera (Lepidoptera: Noctuidae) from Russia. *Journal of Economic Entomology* **88**(1):77–80

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P and Przeworski M (2012) Revisiting an old riddle: What determines genetic diversity levels within species? *PLoS Biol* **10**(9):e1001388

Leite NA, Alves-Pereira A, Corrêa AS, Zucchi MI and Omoto C (2014) Demographics and Genetic Variability of the New World Bollworm (*Helicoverpa* zea) and the Old World Bollworm (*textitHelicoverpa armigera*) in Brazil. *PLOS ONE* **9**(11):e113286

Li X, Schuler MA and Berenbaum MR (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. *Annual Review of Entomology* **52**:231–253

Li X, Zhu B, Gao X and Liang P (2016) Over-expression of UDP–glycosyltransferase gene UGT2B17 is involved in chlorantraniliprole resistance in Plutella xylostella (L.). *Pest Management Science* **73**

Liang J, Zhang L, Xiang Z and He N (2010) Expression profile of cuticular genes of silkworm, Bombyx mori. *BMC Genomics* **11**:173

Liao Y, Smyth GK and Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41**(10):e108

Liao Y, Smyth GK and Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)* **30**(7):923–930

Librado P and Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**(11):1451–1452

Lin Q, Jin F, Hu Z, Chen H, Yin F, Li Z, Dong X, Zhang D, Ren S and Feng X (2013) Transcriptome Analysis of Chlorantraniliprole Resistance Development in the Diamondback Moth Plutella xylostella. *PLOS ONE* **8**(8):e72314

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RRH, Barron M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, Mackey AJ, Munidasa M, Muzny DM, Nazareth L, Newsham I, Perales L, Pu LL, Qu C, Ramia M, Reid JG, Rollmann SM, Rozas J, Saada N, Turlapati L, Worley KC, Wu YQ, Yamamoto A, Zhu Y, Bergman CM, Thornton KR, Mittelman D and Gibbs RA (2012) The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**(7384):173–178

Mahon RJ, Olsen KM and Downes S (2008) Isolations of *Cry*2Ab resistance in Australian populations of *Helicoverpa* armigera (Lepidoptera: Noctuidae) are allelic. *Journal of Economic Entomology* **101**(3):909–914

Mallet J, Korman A, Heckel DG and King P (1993) Biochemical genetics of *Heliothis* and *Helicoverpa* (Lepidoptera: Noctuidae) and evidence for a founder event in *Helicoverpa* zea. *Annals of the Entomological Society of America* **86**(2):189–197

Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, Ferguson HM, Eskin E and Lanzaro GC (2014) Diversity, differentiation, and linkage disequilibrium: prospects for association mapping in the malaria vector Anopheles arabiensis. *G3 (Bethesda)* **4**(1):121–131

Marsden CD, Lee Y, Nieman CC, Sanford MR, Dinis J, Martins C, Rodrigues A, Cornel AJ and Lanzaro GC (2011) Asymmetric introgression between the M and S forms of the malaria vector, Anopheles gambiae, maintains divergence despite extensive hybridization. *Molecular Ecology* **20**(23):4983–4994

Martinez-Torres D, Devonshire AL and Williamson MS (1997) Molecular studies of knockdown resistance to pyrethroids: cloning of domain II sodium channel gene sequences from insects. *Pesticide Science* **51**(3):265–270

Mastrangelo T, Paulo DF, Bergamo LW, Morais EGF, Silva M, Bezerra-Silva G and Azeredo-Espin AML (2014) Detection and genetic diversity of a heliothine invader (Lepidoptera: Noctuidae) from north and northeast of Brazil. *Journal of Economic Entomology* **107**(3):970–980

Matthews M (1991) *Classification of the Heliothinae*. Natural Resources Institute. ISBN 978-0-85954-292-0

Matthews M (1999) *Heliothine moths of Australia: A guide to pest bollworms and related noctuid groups*. Monographs on Australian Lepidoptera ; vol. 7. CSIRO Publishing, Melbourne. ISBN 0-643-06305-6

McCaffery AR, Head DJ, Jianguo T, Dubbeldam AA, Subramaniam VR and Callaghan A (1997) Nerve insensitivity resistance to pyrethroids in Heliothine Lepidoptera. *Pesticide Science* **51**(3):315–320

McCarthy DJ, Chen Y and Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**(10):4288–4297

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**(9):1297–1303

McKenzie JA (2000) The character or the variation: the genetic analysis of the insecticide-resistance phenotype. *Bulletin of entomological research* **90**(01):3–7

McKenzie JA and Batterham P (1994) The genetic, molecular and phenotypic consequences of selection for insecticide resistance. *Trends in Ecology & Evolution* **9**(5):166–169

Mestres J (2005) Structure conservation in cytochromes P450. *Proteins: Structure, Function, and Bioinformatics* **58**(3):596–609

Misra J, Horner M, Lam G and Thummel C (2011) Transcriptional regulation of xenobiotic detoxification in Drosophila. *Genes & Development* **25**(17):1796–1806

Mitchell A and Gopurenko D (2016) DNA Barcoding the Heliothinae (Lepidoptera: Noctuidae) of Australia and Utility of DNA Barcodes for Pest Identification in Helicoverpa and Relatives. *PLOS ONE* **11**(8):e0160895

Mitter C, Poole RW and Matthews M (1993) *Biosystematics of the Heliothinae (Lepidoptera: Noctuidae)*

Morlais I and Severson DW (2003) Intraspecific DNA variation in nuclear genes of the mosquito *Aedes aegypti*. *Insect Molecular Biology* **12**(6):631–639

Murúa MG, Scalora FS, Navarro FR, Cazado LE, Casmuz A, Villagrán ME, Lobos E and Gastaminza G (2014) First Record of *Helicoverpa armigera* (Lepidoptera: Noctuidae) in Argentina. *Florida Entomologist* **97**(2):854–856

Nansen C, Baissac O, Nansen M, Powis K and Baker G (2016) Behavioral Avoidance - Will Physiological Insecticide Resistance Level of Insect Strains Affect Their Oviposition and Movement Responses? *PLOS ONE* **11**(3):e0149994

Nibouche S, Bues R, Toubon JF and Poitout S (1998) Allozyme polymorphism in the cotton bollworm *Helicoverpa* armigera (Lepidoptera: Noctuidae): comparison of African and European populations. *Heredity* **80**(4):438–445

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review Of Genetics* **39**(1):197–218

Noppun V, Saito T and Miyata T (1989) Cuticular penetration of S-fenvalerate in fenvalerate-resistant and susceptible strains of the diamondback moth, Plutella xylostella (L.). *Pesticide Biochemistry and Physiology* **33**(1):83–87

Norris LC, Main BJ, Lee Y, Collier TC, Fofana A, Cornel AJ and Lanzaro GC (2015) Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proceedings of the National Academy of Sciences* **112**(3):815–820

Oakeshott JG, Farnsworth CA, East PD, Scott C, Han Y, Wu Y and Russell RJ (2013) How many genetic options for evolving insecticide resistance in heliothine and spodopteran pests? *Pest Manag. Sci.* **69**(8):889–896

Ometto L, Stephan W and De Lorenzo D (2005) Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila* melanogaster introns and intergenic regions. *Genetics* **169**(3):1521–1527

Ortelli F, Rossiter LC, Vontas J, Ranson H and Hemingway J (2003) Heterologous expression of four glutathione transferase genes genetically linked to a major insecticide-resistance locus from the malaria vector Anopheles gambiae. *The Biochemical Journal* **373**(Pt 3):957–963

O'Loughlin SM, Magesa S, Mbogo C, Mosha F, Midega J, Lomas S and Burt A (2014) Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. *Mol Biol Evol* **31**(4):889–902

Pearce SL, Clarke DF, East PD, Elfekih S, Gordon KHJ, Jermiin LS, McGaughran A, Oakeshott JG, Papanikolaou A, Perera OP, Rane RV, Richards S, Tay WT, Walsh TK, Anderson A, Anderson CJ, Asgari S, Board PG, Bretschneider A, Campbell PM, Chertemps T, Christeller JT, Coppin CW, Downes SJ, Duan G, Farnsworth CA, Good RT, Han LB, Han YC, Hatje K, Horne I, Huang YP, Hughes DST, Jacquin-Joly E, James W, Jhangiani S, Kollmar M, Kuwar SS, Li S, Liu NY, Maibeche MT, Miller JR, Montagne N, Perry T, Qu J, Song SV, Sutton GG, Vogel H, Walenz BP, Xu W, Zhang HJ, Zou Z, Batterham P, Edwards OR, Feyereisen R, Gibbs RA, Heckel DG, McGrath A, Robin C, Scherer SE, Worley KC and Wu YD (2017) Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive Helicoverpa pest species. *BMC Biology* **15**(1):63

Pedra JHF, McIntyre LM, Scharf ME and Pittendrigh BR (2004) Genome-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT)-resistant Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **101**(18):7034–7039

Pedrini N, Mijailovsky SJ, Girotti JR, Stariolo R, Cardozo RM, Gentile A and Juárez MP (2009) Control of Pyrethroid-Resistant Chagas Disease Vectors with Entomopathogenic Fungi. *PLOS Neglected Tropical Diseases* **3**(5):e434

Pittendrigh B, Aronstein K, Zinkovsky E, Andreev O, Campbell B, Daly J, Trowell S and Ffrench-Constant RH (1997) Cytochrome P450 genes from *Helicoverpa armigera*: Expression in a pyrethroid-susceptible and -resistant strain. *Insect Biochemistry and Molecular Biology* **27**(6):507–512

Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, Baglione V, Unneberg P, Wikelski M, Grabherr MG and Wolf JBW (2014) The genomic

landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**(6190):1410–1414

Pogue MG (2013) Revised status of Chloridea Duncan and (Westwood), 1841, for the Heliothis virescens species group (Lepidoptera: Noctuidae: Heliothinae) based on morphology and three genes. *Systematic Entomology* **38**(3):523–542

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ and Langley CH (2012) Population genomics of sub-Saharan *Drosophila* melanogaster: African diversity and non-African admixture. *PLoS Genet* **8**(12):e1003080

Poulos TL and Johnson EF (2005) Structures of Cytochrome P450 Enzymes. *SpringerLink* pp. 87–114

Prapanthadara LA, Hemingway J and Ketterman AJ (1993) Partial Purification and Characterization of Glutathione S-Transferases Involved in DDT Resistance from the Mosquito Anopheles gambiae. *Pesticide Biochemistry and Physiology* **47**(2):119–133

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**(2):945–959

Puinean AM, Foster SP, Oliphant L, Denholm I, Field LM, Millar NS, Williamson MS and Bass C (2010) Amplification of a Cytochrome P450 Gene Is Associated with Resistance to Neonicotinoid Insecticides in the Aphid Myzus persicae. *PLOS Genetics* **6**(6):e1000999

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and Sham PC (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* **81**(3):559–575

Qiu X, Sun W, McDonnell CM, Li-Byarlay H, Steele LD, Wu J, Xie J, Muir WM and Pittendrigh BR (2013) Genome-wide analysis of genes associated with

moderate and high DDT resistance in *Drosophila melanogaster*. *Pest Management Science* **69**(8):930–937

Qiu Y, Tittiger C, Wicker-Thomas C, Goff GL, Young S, Wajnberg E, Fricaux T, Taquet N, Blomquist GJ and Feyereisen R (2012) An insect-specific P450 oxidative decarbonylase for cuticular hydrocarbon biosynthesis. *Proceedings of the National Academy of Sciences* **109**(37):14858–14863

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

Ranasinghe C, Campbell B and Hobbs AA (1998) Over-expression of cytochrome P450 CYP6B7 mRNA and pyrethroid resistance in Australian populations of *Helicoverpa armigera* (Hübner). *Pesticide Science* **54**(3):195–202

Ranasinghe C and Hobbs AA (1998) Isolation and characterization of two cytochrome P450 cDNA clones for CYP6B6 and CYP6B7 from *Helicoverpa armigera* (Hubner): possible involvement of CYP6B7 in pyrethroid resistance. *Insect Biochemistry and Molecular Biology* **28**(8):571–580

Rane RV, Walsh TK, Pearce SL, Jermiin LS, Gordon KH, Richards S and Oakeshott JG (2016) Are feeding preferences and insecticide resistance associated with the size of detoxifying enzyme families in insect herbivores? *Current Opinion in Insect Science* **13**(Supplement C):70–76

Rasool A, Joußen N, Lorenz S, Ellinger R, Schneider B, Khan SA, Ashfaq M and Heckel DG (2014) An independent occurrence of the chimeric P450 enzyme CYP337B3 of *Helicoverpa armigera* confers cypermethrin resistance in Pakistan. *Insect Biochemistry and Molecular Biology* **53**:54–65

Rauschenbach IY, Gruntenko NE, Khlebodarova TM, Mazurov MM, Grenback LG, Sukhanova MJ, Shumnaja LV, Zakharov IK and Hammock BD (1996) The role of the degradation system of the juvenile hormone in the reproduction of Drosophila under stress. *Journal of Insect Physiology* **42**(8):735–742

Rašić G, Filipović I, Weeks AR and Hoffmann AA (2014) Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, \textAedes aegypti. *BMC Genomics* **15**:275

Reinhardt JA, Kolaczkowski B, Jones CD, Begun DJ and Kern AD (2014) Parallel geographic variation in *Drosophila melanogaster*. *Genetics* **197**(1):361–373

Ren X, Han Z and Wang Y (2002) Mechanisms of monocrotophos resistance in cotton bollworm, *Helicoverpa armigera* (Hübner). *Archives of Insect Biochemistry and Physiology* **51**(3):103–110

Robinson MD, McCarthy DJ and Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**(1):139–140

Roelofs D, Morgan J and Stürzenbaum S (2010) The significance of genome-wide transcriptional regulation in the evolution of stress tolerance. *Evolutionary Ecology* **24**(3):527–539

Rohland N and Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* **22**(5):939–946

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, Loire E, Lourenco JM, Nabholz B, Roux C, Tsagkogeorga G, Weber AaT, Weinert LA, Belkhir K, Bierne N, Glémin S and Galtier N (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**(7526):261–263

Rosenberg NA and Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3**(5):380–390

Roush RT and Tabashnik BE (editors) (1991) *Pesticide Resistance in Arthropods*. Springer US, Boston, MA. ISBN 978-1-4684-6431-3 978-1-4684-6429-0

Rozas J, Sanchez-DelBarrio JC, Messeguer X and Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**(18):2496–2497

Russell RJ, Claudianos C, Campbell PM, Horne I, Sutherland TD and Oakeshott JG (2004) Two major classes of target site insensitivity mutations confer resistance to organophosphate and carbamate insecticides. *Pesticide Biochemistry and Physiology* **79**(3):84–93

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP and Hartl DL (2014) Positive selection drives faster-Z evolution in silkmoths. *Evolution* **68**(8):2331–2342

Sanchez-Gracia A and Rozas J (2007) Unusual pattern of nucleotide sequence variation at the *OS*-E and *OS*-F genomic regions of *Drosophila* simulans. *Genetics* **175**(4):1923

Sawicki RM (1970) Interaction between the factor delaying penetration of insecticides and the desethylation mechanism of resistance in organophosphorus-resistant houseflies. *Pesticide Science* **1**(3):84–87

Sawicki RM and Lord KA (1970) Some properties of a mechanism delaying penetration of insecticides into houseflies. *Pesticide Science* **1**(5):213–217

Schlenke TA and Begun DJ (2004) Strong selective sweep associated with a transposon insertion in *Drosophila* simulans. *Proceedings of the National Academy of Sciences of the United States of America* **101**(6):1626–1631

Scott KD, Lange CL, Scott LJ and Graham GC (2004) Isolation and characterization of microsatellite loci from *Helicoverpa armigera* Hubner (Lepidoptera: Noctuidae). *Mol Ecol Notes* **4**(2):204–205

Scott KD, Wilkinson KS, Merritt MA, Scott LJ, Lange CL, Schutze MK, Kent JK, Merritt DJ, Grundy PR and Graham GC (2003) Genetic shifts in *Helicoverpa armigera* Hübner (Lepidoptera: Noctuidae) over a year in the Dawson/Callide Valleys. *Aust. J. Agric. Res.* **54**(8):739–744

Seehuus SC, Norberg K, Gimsa U, Krekling T and Amdam GV (2006) Reproductive protein protects functionally sterile honey bee workers from oxidative stress. *Proceedings of the National Academy of Sciences* **103**(4):962–967

Silva AX, Jander G, Samaniego H, Ramsey JS and Figueroa CC (2012) Insecticide Resistance Mechanisms in the Green Peach Aphid Myzus persicae (Hemiptera: Aphididae) I: A Transcriptomic Survey. *PLOS ONE* **7**(6):e36366

Slater G and Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**(1):31

Slatkin M (2008) Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**(6):477–485

Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, Nachman MW and Kohn MH (2011) Adaptive introgression of anticoagulant rodent poison resistance by hybridization between Old World mice. *Current biology : CB* **21**(15):1296–1301

Sosa-Gómez DR, Specht A, Paula-Moraes SV, Lopes-Lima A, Yano SAC, Micheli A, Morais EGF, Gallo P, Pereira PRVS, Salvadori JR, Botton M, Zenker MM and Azevedo-Filho WS (2016) Timeline and geographical distribution of *Helicoverpa armigera* (Hübner) (Lepidoptera, Noctuidae: Heliothinae) in Brazil. *Revista Brasileira de Entomologia* **60**(1):101–104

Srinivas R, Udikeri SS, Jayalakshmi SK and Sreeramulu K (2004) Identification of factors responsible for insecticide resistance in *Helicoverpa armigera*. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **137**(3):261–269

Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA and Tautz D (2012) Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse *(Mus* musculus). *PLoS Genet* **8**(8):e1002891

Stokes NH, McKechnie SW and Forrester NW (1997) Multiple allelic variation in a sodium channel gene from populations of Australian *Helicoverpa* armigera (Hübner) (Lepidoptera: Noctuidae) detected via temperature gradient gel electrophoresis. *Australian Journal of Entomology* **36**(2):191–196

Sugiyama H, Moriyama H and Shigematsu H (1981) Specific accumulation of p-hydroxyphenyllactic acid as a metabolite of tyrosine metabolism in the haemolymph

of the larva of *Bombyx* mori L.: Lepidoptera: Bombycidae: Intoxicated with MTMC (m-tolyl-N-methylcarbamate). *Applied entomology and zoology* **16**(4):472–476

Tabashnik BE, Gould F and Carrière Y (2004) Delaying evolution of insect resistance to transgenic crops by decreasing dominance and heritability. *Journal of Evolutionary Biology* **17**(4):904–912

Tajima F (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**(3):585–595

Tauchman SJ, Lorch JM, Orth AP and Goodman WG (2007) Effects of stress on the hemolymph juvenile hormone binding protein titers of Manduca sexta. *Insect Biochemistry and Molecular Biology* **37**(8):847–854

Tay W, Behere G, Batterham P and Heckel D (2010) Generation of microsatellite repeat families by RTE retrotransposons in lepidopteran genomes. *BMC Evolutionary Biology* **10**(1):144

Tay W, Behere G, Heckel D, Lee S and Batterham P (2008) Exon-primed intron-crossing (EPIC) PCR markers of *Helicoverpa* armigera (Lepidoptera: Noctuidae). *Bulletin of Entomological Research* **98**(05):509–518

Tay WT, Mahon RJ, Heckel DG, Walsh TK, Downes S, James WJ, Lee SF, Reineke A, Williams AK and Gordon KHJ (2015) Insect Resistance to Bacillus thuringiensis Toxin Cry2Ab Is Conferred by Mutations in an ABC Transporter Subfamily A Protein. *PLOS Genetics* **11**(11):e1005534

Tay WT, Soria MF, Walsh T, Thomazoni D, Silvie P, Behere GT, Anderson C and Downes S (2013) A brave new world for an Old World pest: *Helicoverpa* armigera (Lepidoptera: Noctuidae) in Brazil. *PLoS ONE* **8**(11):e80134

Teese MG, Farnsworth CA, Li Y, Coppin CW, Devonshire AL, Scott C, East P, Russell RJ and Oakeshott JG (2013) Heterologous Expression and Biochemical Characterisation of Fourteen Esterases from Helicoverpa armigera. *PLoS ONE* **8**(6)

Tian K, Liu D, Yuan Y, Li M and Qiu X (2017) CYP6B6 is involved in esfenvalerate detoxification in the polyphagous lepidopteran pest, *Helicoverpa armigera*. *Pesticide Biochemistry and Physiology* **138**(Supplement C):51–56

Togawa T, Dunn WA, Emmons AC, Nagao J and Willis JH (2008) Developmental expression patterns of cuticular protein genes with the R&R Consensus from Anopheles gambiae. *Insect Biochemistry and Molecular Biology* **38**(5):508–519

Toé KH, N'Falé S, Dabiré RK, Ranson H and Jones CM (2015) The recent escalation in strength of pyrethroid resistance in Anopheles coluzzi in West Africa is linked to increased expression of multiple gene families. *BMC Genomics* **16**(1)

Vicoso B and Charlesworth B (2006) Evolution on the X chromosome: unusual patterns and processes. *Nature Reviews Genetics* **7**(8):645–653

Vijaykumar, Bashasab F, Krishnareddy KB, Kuruvinashetti MS and Patil BV (2007) Mating Compatibility Among *Helicoverpa armigera* (Lepidoptera: Noctuidae) Occurring on Selected Host Plants and Bt Cotton Survivors. *Journal of Economic Entomology* **100**(3):903–908

Vontas J, Blass C, Koutsos AC, David JP, Kafatos FC, Louis C, Hemingway J, Christophides GK and Ranson H (2005) Gene expression in insecticide resistant and susceptible Anopheles gambiae strains constitutively or after insecticide exposure. *Insect Molecular Biology* **14**(5):509–521

Vontas JG, Small GJ and Hemingway J (2001) Glutathione S-transferases as antioxidant defence agents confer pyrethroid resistance in Nilaparvata lugens. *The Biochemical Journal* **357**(Pt 1):65–72

Wallar BJ and Alberts AS (2003) The formins: active scaffolds that remodel the cytoskeleton. *Trends in Cell Biology* **13**(8):435–446

Wang C and Dong J (2001) Interspecific hybridization of *Helicoverpa* armigera and *H.* assulta (Lepidoptera: Noctuidae). *Chinese Science Bulletin* **46**(6):489–491

Wang C, Scharf ME and Bennett GW (2004) Behavioral and Physiological Resistance of the German Cockroach to Gel Baits (Blattodea: Blattellidae). *Journal of Economic Entomology* **97**(6):2067–2072

Wang XP and Hobbs AA (1995) Isolation and sequence analysis of a cDNA clone for a pyrethroid inducible cytochrome P450 from *Helicoverpa armigera*. *Insect Biochemistry and Molecular Biology* **25**(9):1001–1009

Wee CW, Lee SF, Robin C and Heckel DG (2008) Identification of candidate genes for fenvalerate resistance in *Helicoverpa* armigera using cDNA-AFLP. *Insect Molecular Biology* **17**(4):351–360

Weeks A, Endersby N, Lange C, Lowe A, Zalucki M and Hoffmann A (2010) Genetic Variation Among Helicoverpa armigera Populations as Assessed by Microsatellites: A Cautionary Tale About Accurate Allele Scoring. *Bulletin of Entomological Research* **100**(04):445–450

Weill M, Chandre F, Brengues C, Manguin S, Akogbeto M, Pasteur N, Guillet P and Raymond M (2000) The kdr mutation occurs in the Mopti form of Anopheles gambiaes.s. through introgression. *Insect Molecular Biology* **9**(5):451–455

Wickham H (2009) ggplot2: Elegant graphics for data analysis. *New York, USA* **907**

Wilding C, Weetman D, Steen K and Donnelly M (2009) High, clustered, nucleotide diversity in the genome of *Anopheles* gambiae revealed through pooled-template sequencing: implications for high-throughput genotyping protocols. *BMC Genomics* **10**(1):320

Willingham AT and Keil T (2004) A tissue specific cytochrome P450 required for the structure and function of Drosophila sensory organs. *Mechanisms of Development* **121**(10):1289–1297

Willoughby L, Chung H, Lumb C, Robin C, Batterham P and Daborn PJ (2006) A comparison of *Drosophila melanogaster* detoxification gene induction responses for six insecticides, caffeine and phenobarbital. *Insect Biochemistry and Molecular Biology* **36**(12):934–942

Wilson AGL (1974) Resistance of Heliothis armigera to Insecticides in the Ord Irrigation Area, North Western Australia. *Journal of Economic Entomology* **67**(2):256–258

Wondji CS, Hemingway J and Ranson H (2007) Identification and analysis of Single Nucleotide Polymorphisms (SNPs) in the mosquito Anopheles funestus, malaria vector. *BMC Genomics* **8**(1):1–13

Wood OR, Hanrahan S, Coetzee M, Koekemoer LL and Brooke BD (2010) Cuticle thickening associated with pyrethroid resistance in the major malaria vector Anopheles funestus. *Parasites & Vectors* **3**:67

Wu S, Yang Y, Yuan G, Campbell PM, Teese MG, Russell RJ, Oakeshott JG and Wu Y (2011) Overexpressed esterases in a fenvalerate resistant strain of the cotton bollworm, *Helicoverpa armigera*. *Insect Biochemistry and Molecular Biology* **41**(1):14–21

Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z, Dai F, Li Y, Cheng D, Li R, Cheng T, Jiang T, Becquet C, Xu X, Liu C, Zha X, Fan W, Lin Y, Shen Y, Jiang L, Jensen J, Hellmann I, Tang S, Zhao P, Xu H, Yu C, Zhang G, Li J, Cao J, Liu S, He N, Zhou Y, Liu H, Zhao J, Ye C, Du Z, Pan G, Zhao A, Shao H, Zeng W, Wu P, Li C, Pan M, Li J, Yin X, Li D, Wang J, Zheng H, Wang W, Zhang X, Li S, Yang H, Lu C, Nielsen R, Zhou Z, Wang J, Xiang Z and Wang J (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm *(Bombyx)*. *Science* **326**(5951):433–436

Xiang M, Zhang L, Lu Y, Tang Q, Liang P, Shi X, Song D and Gao X (2017) A P-glycoprotein gene serves as a component of the protective mechanisms against 2-tridecanone and abamectin in *Helicoverpa armigera*. *Gene* **627**:63–71

Xiao Y, Liu K, Zhang D, Gong L, He F, Soberón M, Bravo A, Tabashnik BE and Wu K (2016) Resistance to Bacillus thuringiensis Mediated by an ABC Transporter Mutation Increases Susceptibility to Toxins from Other Bacteria in an Invasive Insect. *PLOS Pathogens* **12**(2):e1005450

Xiao Y, Zhang T, Liu C, Heckel DG, Li X, Tabashnik BE and Wu K (2014) Mis-splicing of the ABCC2 gene linked with Bt toxin resistance in *textitHelicoverpa armigera*. Scientific Reports* **4**:srep06184

Xu C, Li CYT and Kong ANT (2005) Induction of phase I, II and III drug metabolism/transport by xenobiotics. *Archives of Pharmacal Research* **28**(3):249

Xu L, Li D, Qin J, Zhao W and Qiu L (2016) Over-expression of multiple cytochrome P450 genes in fenvalerate-resistant field strains of *Helicoverpa armigera* from north of China. *Pesticide Biochemistry and Physiology* **132**(Supplement C):53–58

Yang Y, Li Y and Wu Y (2013) Current Status of Insecticide Resistance in *Helicoverpa armigera* After 15 Years of Bt Cotton Planting in China. *Journal of Economic Entomology* **106**(1):375–381

Yang Y, Wu Y, Chen S, Devine G, Denholm I, Jewess P and Moores G (2004) The involvement of microsomal oxidases in pyrethroid resistance in *Helicoverpa armigera* from Asia. *Insect Biochemistry and Molecular Biology* **34**(8):763–773

Yang Y, Yue L, Chen S and Wu Y (2008) Functional expression of *Helicoverpa armigera* CYP9A12 and CYP9A14 in Saccharomyces cerevisiae. *Pesticide Biochemistry and Physiology* **92**(2):101–105

Yasukochi Y, Ashakumary LA, Baba K, Yoshido A and Sahara K (2006) A second-generation integrated map of the silkworm reveals synteny and conserved gene order between lepidopteran insects. *Genetics* **173**(3):1319–1328

Young SJ, Gunning RV and Moores GD (2005) The effect of piperonyl butoxide on pyrethroid-resistance-associated esterases in *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae). *Pest Management Science* **61**(4):397–401

Zalucki M, Daglish G, Firempong S and Twine P (1986) The biology and ecology of *Heliothis* armigera (Hubner) and *Heliothis* punctigera Wallengren (Lepidoptera, Noctuidae) in Australia: What do we know? *Aust. J. Zool.* **34**(6):779–814

Zalucki M and Furlong M (2017) Behavior as a mechanism of insecticide resistance: evaluation of the evidence. *Current Opinion in Insect Science* **21**(Supplement C):19–25

Zhang DX (2004) Lepidopteran microsatellite DNA: Redundant but promising. *Trends in Ecology & Evolution* **19**(10):507–509

Zhang H, Tang T, Cheng Y, Shui R, Zhang W and Qiu L (2010) Cloning and expression of cytochrome P450 CYP6B7 in fenvalerate-resistant and susceptible *Helicoverpa armigera* (Hübner) from China. *Journal of Applied Entomology* **134**(9-10):754–761

Zhao G, Rose RL, Hodgson E and Roe RM (1996) Biochemical Mechanisms and Diagnostic Microassays for Pyrethroid, Carbamate, and Organophosphate Insecticide Resistance/Cross-Resistance in the Tobacco Budworm,Heliothis virescens. *Pesticide Biochemistry and Physiology* **56**(3):183–195

Zhou X, Faktor O, Applebaum SW and Coll M (2000) Population structure of the pestiferous moth *textitHelicoverpa armigera* in the Eastern Mediterranean using RAPD analysis. *Heredity* **85**(3):251–256

Zhou X, Ma C, Li M, Sheng C, Liu H and Qiu X (2010a) CYP9A12 and CYP9A17 in the cotton bollworm, *textitHelicoverpa armigera*: Sequence similarity, expression profile and xenobiotic response. *Pest Management Science* **66**(1):65–73

Zhou X, Sheng C, Li M, Wan H, Liu D and Qiu X (2010b) Expression responses of nine cytochrome P450 genes to xenobiotics in the cotton bollworm *Helicoverpa armigera*. *Pesticide Biochemistry and Physiology* **97**(3):209–213

Zhu L, Tatsuke T, Li Z, Mon H, Xu J, Lee JM and Kusakabe T (2012) Molecular cloning of BmTUDOR-SN and analysis of its role in the RNAi pathway in the silkworm, Bombyx mori (Lepidoptera: Bombycidae). *Applied Entomology and Zoology* **47**(3):207–215

Zhu L, Tatsuke T, Mon H, Li Z, Xu J, Lee JM and Kusakabe T (2013) Characterization of Tudor-sn-containing granules in the silkworm, Bombyx mori. *Insect biochemistry and molecular biology* **43**(8):664–674

Appendices

Appendix A

Protocols for RNA-Seq library preparation (Chris Coppin)

mRNA isolation NEB Oligo d(T)$_{25}$ Magnetic Beads

Version 3

Denaturation

1. Transfer 50 µL aliquots of total RNA to an **Axygen V-bottom assay plate**. Cover samples with a foil seal.
2. Heat plate at 65°C for 2 min then immediately chill on ice.

I. Bead preparation

1. Wash enough **Oligo d(T)$_{25}$ Magnetic Beads** for the number of samples to be processed:
   a. Add 15 µL of **Oligo d(T)$_{25}$ Magnetic Beads** per sample (360 µL for 24 samples) to a 1.5 mL tube. Place tube on magnetic stand and allow beads to separate. Remove and discard supernatant.
   b. Wash beads by adding 50 µL of **Binding Buffer** per sample and resuspending. Place tube on magnetic stand and allow beads to separate. Remove and discard supernatant.
   c. Repeat step b. *i.e* Wash twice
   d. Resuspend washed beads in 50 µL of **Binding Buffer** per sample.

II. mRNA isolation (Round 1)

1. Remove foil from the **denatured total RNA** samples and add 50 µL of **washed Oligo d(T)$_{25}$ Magnetic Beads**. Re-seal the samples and incubate at room temperature for 15 min on a Titramax plate shaker at 1200 rpm.
2. Remove foil and insert magnetic separator with cover plate. Allow magnetic beads to bind to the magnets (~2 min).
3. Transfer magnetic beads (attached to the separator) to a fresh well containing 120 µL of **Washing Buffer** and soak briefly.
4. Repeat step 3. *i.e* Wash twice
5. Transfer magnetic beads to a fresh well containing 50 µL of **Elution Buffer**. Remove magnetic separator and move cover plate around in the buffer to resuspend the beads. When the beads have been sufficiently resuspended (there are few beads still attached to the cover plate) remove the cover plate cover and return to the magnetic separator for re-use. Cover the samples with a fresh foil seal.
6. Incubate plate at 80°C for 2 min to elute the mRNA. Immediately remove foil and insert magnetic separator with cover plate and allow beads to separate.
7. Remove the beads and re-seal the plate – well contains eluted mRNA. Do not discard the beads.

III. Bead regeneration

1. Wash the used beads for second round mRNA isolation:
   a. Transfer magnetic beads to a fresh well containing 120 µL of **nuclease-free water**, remove magnetic separator and resuspend the beads.
   b. Re-insert the magnetic separator into the cover plate and allow beads to separate.
   c. Repeat steps a-b. *i.e* Wash twice

IV. mRNA isolation (Round 2)

1. Remove foil from the **eluted mRNA** samples and add 50 µL of **Binding Buffer**.
2. Transfer the regenerated magnetic beads back into corresponding mRNA sample well. Remove magnetic separator, resuspend the beads, then remove the cover plate cover and return to the magnetic separator for re-use. Re-seal the samples and incubate at room temperature for 15 min on a Titramax plate shaker at 1200 rpm.
3. Remove foil and insert magnetic separator with cover plate. Allow magnetic beads to bind to the magnets.
4. Transfer magnetic beads to a fresh well containing 120 µL of **Washing Buffer 2** and soak briefly.
5. Repeat step 4. *i.e* Wash twice
6. Transfer magnetic beads to a fresh well containing 10 µL of **Elution Buffer**. Remove magnetic separator, resuspend the beads, then remove the cover plate cover and return to the magnetic separator for re-use. Cover the samples with a fresh foil seal.
7. Incubate tubes at 80°C for 2 min to elute the mRNA. Immediately place tube on magnetic stand and allow beads to separate.
8. Transfer the supernatant containing the eluted mRNA to a fresh PCR tube and store on ice.

V. Sample division

1. Aliquot 3 µL into each of two PCR tubes - to be archived at -80°C as backup. The 1 µL of the remaining volume may be used for quantification and quality control and the rest stored at -80°C in preparation of first strand synthesis.

Buffers: (all Nuclease-free reagents – either certified or baked).

Binding Buffer (20 mM TrisHCl pH 7.5, 1 M LiCl, 2 mM EDTA)

1 M Tris-HCl pH 7.5	1 mL
10 M LiCl	5 mL
500 mM EDTA	0.2 mL
Nuclease-free water	to 50 mL in a Nuclease-free tube

Elution Buffer (10 mM TrisHCl pH 7.5)

1 M Tris-HCl pH 7.5	0.5 mL
Nuclease-free water	to 50 mL in a Nuclease-free tube

Wash Buffer (10 mM TrisHCl pH 7.5, 150 mM LiCl, 2 mM EDTA)

1 M Tris-HCl pH 7.5	0.5 mL
10 M LiCl	0.75 mL
500 mM EDTA	0.2 mL
Nuclease-free water	to 50 mL in a Nuclease-free tube

Wash Buffer 2 (10 mM TrisHCl pH 7.5, 150 mM LiCl)

1 M Tris-HCl pH 7.5	0.5 mL
10 M LiCl	0.75 mL
Nuclease-free water	to 50 mL in a Nuclease-free tube

First strand cDNA synthesis with Enzscript (M-MLV RNase H minus Reverse Transcriptase) Version 1

Fragmentation/random priming

1. Make up a appropriate volume of **Fragmentation Buffer** and dispense 4 µL into PCR strip tubes:

	Per reaction	8 reactions	24 reactions	32 reactions	54 reactions
5× Enzscript Reverse transcriptase buffer	1 µL	10 µL	26 µL	34 µL	56 µL
10 µM RT-Hex primer	1 µL	10 µL	26 µL	34 µL	56 µL
Nuclease-free water	2 µL	20 µL	52 µL	68 µL	112 µL

2. Add 1 µL of mRNA, flick mix and spin down.
3. Incubate samples at 85°C for exactly 5 min and then transfer immediately onto ice.

I. Reverse transcription/template switching

1. Make up a appropriate volume of **RT Reaction Mix** and add 5 µL to each fragmented mRNA:

	Per reaction	8 reactions	24 reactions	32 reactions	54 reactions
Nuclease-free water	0.75µL	7.5µL	19.5 µL	25.5 µL	42 µL
10 mM dNTPs	1 µL	10 µL	26 µL	34 µL	56 µL
5× Enzscript Reverse transcriptase buffer	1 µL	10 µL	26 µL	34 µL	56 µL
100 mM DTT	0.5 µL	5 µL	13 µL	17 µL	28 µL
Enzymatics RNase inhibitor	0.25 µL	2.5 µL	6.5 µL	8.5 µL	14 µL
10 µM Bio_TS_RNA primer	1 µL	10 µL	26 µL	34 µL	56 µL
Enzymatics Enzscript	0.5 ul	5 µL	13 µL	17 µL	28 µL

2. Flick mix samples, spin down and incubate as follows:

	Temperature	Duration
Lid	45°C	—
Block	4°C	Hold → Start
	25°C	10 min
	42°C	90 min
	72°C	10 min
	4°C	Hold

II. Sample cleanup (Round 1)

1. Add 9.5 µL of **SeraPure RNA beads** to each completed reverse transcriptase reaction. Flick mix samples and spin down briefly.
2. Incubate samples at room temperature for 8 min to allow the mRNA/cDNA hybrid to bind to the beads.
3. Transfer the samples to a V-bottom assay plate then insert magnetic separator with cover plate. Allow magnetic beads to bind to the magnets (~2 min).
4. Transfer magnetic beads (attached to the separator) to a fresh well containing 200 µL of fresh **80 % ethanol**.
5. Repeat step 4. *i.e* Wash twice
6. Remove the magnetic separator with beads attached, invert and allow to air dry for 5 min.
7. Transfer the dry magnetic beads to a fresh well containing 10 µL of **nuclease-free water**, remove magnetic separator from cover plate and move cover plate around in the water to resuspend the beads.
8. Incubate samples at room temperature for 5 min to elute the mRNA/cDNA hybrid, then place magnetic separator back into the cover plate to allow the beads to rebind to the magnets.
9. Remove beads and discard.

III. Sample cleanup (Round 2)

1. Add 9.5 µL of fresh **SeraPure RNA beads** to each sample in the assay plate. Insert a fresh magnetic separator cover plate into the samples and use it to mix the samples.
2. Incubate samples at room temperature for 8 min to allow the mRNA/cDNA hybrid to bind to the beads.
3. Insert magnetic separator into the cove plater. Allow magnetic beads to bind to the magnets (~2 min).
4. Transfer magnetic beads attached to the separator to a fresh well containing 200 µL of fresh **80 % ethanol**.
5. Repeat step 4. *i.e* Wash twice
6. Remove the magnetic separator with beads attached, invert and allow to air dry for 5 min.
7. Transfer the dry magnetic beads to a fresh well containing 20 µL of **nuclease-free water**, remove magnetic separator from cover plate and move cover plate around in the water to resuspend the beads.
8. Incubate samples at room temperature for 5 min to elute the mRNA/cDNA hybrid, then place magnetic separator back into the cover plate to allow the beads to rebind to the magnets.
9. Remove beads and discard. Transfer supernatant to PCR strip tubes and store at -20°C.

Barcoding reactions

Version 1

I. Barcoding

1. Remove 1st strand cDNA product from freezer and thaw thoroughly.
2. Make up a **bulk PCR reaction mix** according to the following volumes per sample:

	Per reaction
5× Phusion buffer	4 µL
10 mM dNTPs	0.4 µL
H_2O	2.4 µL
Phusion polymerase	0.2 µL
	7 µL

3. Dispense 7 µL of **bulk PCR reaction mix** into the bottom of a PCR strip tube on ice for each sample.
4. Dilute 10 µM aliquots of **barcode primers** (10 µL) to 2 µM by the addition of 40 µL of **nuclease-free water**.
5. Dispense 2.5 µL of each **2 µM barcode primer** across the PCR strip tube array containing the bulk PCR reaction mix as indicated below, ensuring that the i5_*#* and i7_*X* are spotted onto opposite sides of the tube to prevent cross-contamination.

		2 µM i5_#	2.5 µL i5_1	2.5 µL i5_2	2.5 µL i5_3	2.5 µL i5_4	2.5 µL i5_5	2.5 µL i5_6	2.5 µL i5_8
2 µM i7_*X*			▼	▼	▼	▼	▼	▼	▼
			1	2	3	4	5	6	7
2.5 µL i7_A	►	A							
2.5 µL i7_B	►	B							
2.5 µL i7_C	►	C							
2.5 µL i7_D	►	D							
2.5 µL i7_E	►	E							
2.5 µL i7_F	►	F							
2.5 µL i7_G	►	G							
2.5 µL i7_H	►	H							

6. Add 8 µL of 1st strand cDNA product to the lid of the tubes and seal the lid.
7. Mark tubes and separate according to calculated optimal cycle number.
8. Process samples with the same optimal cycle number together – invert and flick mix tubes, then pulse spin then thermocycle as follows in a Kyratec thermocycler. Three machines were run simultaneously, samples yet to be thermocycled were kept on ice.

Step	Cycles	Temperature	Duration
1	1	80°C	Hold
		98°C	10 sec
2	*x*	98°C	5 sec
		58°C	10 sec
		72°C	20 sec
3	*1*	72°C	5 min
		15°C	Hold
Lid = 105°			
Volume = 20 µL			
Use predicted liquid temperature for hold times = ON			
Accelerated Thermal Equilibration = ON			

At the completion of all the barcoding PCRs, re-array the samples in their correct order

II. Sample clean up

1. Add 17 µL (0.85 vol)of **SeraPure RNA beads** to each completed barcoded reaction. Flick mix samples and spin down briefly.
2. Incubate samples at room temperature for 8 min to allow the barcoded DNA to bind to the beads.
3. Transfer the samples to a V-bottom assay plate, pulse spin to ~3000 rpm to remove bubbles, then insert magnetic separator with cover plate. Allow magnetic beads to bind to the magnets (~2 min).
4. Transfer magnetic beads (attached to the separator) to a fresh well containing 200 µL of ice cold **70 % ethanol**.
5. Repeat step 4. *i.e* Wash twice
6. Remove the magnetic separator with beads attached, invert and allow to air dry for 5 min.
7. Transfer the dry magnetic beads to a fresh well containing 20 µL of **10 mM Tris pH 8.0**, remove magnetic separator from cover plate and move cover plate around in the water to resuspend the beads.
8. Incubate samples at room temperature for 5 min to elute the barcoded DNA, then place magnetic separator back into the cover plate to allow the beads to rebind to the magnets.
9. Remove beads and discard. Transfer the supernatant to PCR strip tubes and store at -20°C.

Barcoding cycle number optimization by qPCR

Version 1

I. 1st strand cDNA sample dilution

1. Add a 1 µL sample of the 1st strand cDNA to 15 µL of **nuclease-free water** in PCR strip tubes and mix thoroughly.

II. qPCR reaction

1. Set up duplicate qPCR reactions for each diluted sample:

	Per reaction
Bio-Rad SsoFast EvaGreen Supermix	5 µL
5 µM TS_qPCR/5 µM RT_Hex_qPCR primer mix	1 µL
Diluted cDNA template	4 µL

2. Cycle on a Bio-Rad CFX96 qPCR machine using the following protocol:

Step	Cycles	Temperature	Duration
1	1	95°C	45 sec
2	30	95°C	5 sec
		60°C	30 sec

III. Optimal cycle number

1. Calculate the mean Cq value for each diluted sample. The optimal cycle number for 8 µL of *undiluted* 1st strand cDNA is calculated by subtracting 0.4 from the mean Cq and then rounding *down* to the nearest whole number.

Appendix B

List of DE genes that are shared across the UU and SS strains at the 6, 12 and 24-hour timepoints

	Gene	HaOG	LogFC in	
			UU	SS
	At 6 hours (n=100)			
1	CYP9A17-Ha	HaOG200110	4.23	3.44
2	BMORI:frizzled-2-like	HaOG210988	1.93	3.42
3	BMORI:proton-coupled folate transporter-like	HaOG206635	2.63	3.33
4	BMORI:GTPase-activating protein CdGAPr-like	HaOG213929	2.24	2.88
5	CYP341B7-Ha	HaOG200058	1.55	2.66
6	DMELA:A8JR05 Crossveinless c, isoform C	HaOG214463	2.43	2.56
7	BMORI:uncharacterized protein LOC101745090	HaOG210985	1.72	2.17
8	unknown	HaOG202754	3.26	2.17
9	DMELA:Q9VIC9 CG8665	HaOG207522	2.66	2.09
10	BMORI:putative uncharacterized protein DDB-G0277255-like	HaOG209375	1.93	2.07
11	DMELA:Q9VHC5 CG8165	HaOG213149	1.64	2.06
12	DMELA:E1JH44 Cuticular protein 49Ac, isoform D BMORI:cuticular protein RR-1 motif 1 precursor	HaOG215985	1.53	2.00
13	CYP333B3-Ha	HaOG200024	1.61	1.99
14	BMORI:peroxidase-like isoform X1	HaOG213098	2.97	1.98
15	BMORI:solute carrier family 46 member 3-like	HaOG206634	1.98	1.97
16	DMELA:Q8IPB7 CG6206, isoform B	HaOG214600	1.74	1.94
17	DMELA:Q9VMC9 Kynurenine formamidase BMORI:kynurenine formamidase-like	HaOG211924	2.14	1.91
18	DMELA:Q9VPL5 CG11490	HaOG206862	1.87	1.90
19	BMORI:microtubule-associated protein futsch-like	HaOG207771	2.74	1.88
20	HaCCE033	HaOG200147	1.46	1.88
21	DMELA:A8JNN5 CG7112, isoform B BMORI:rab GTPase-activating protein 1-like isoform X1	HaOG213468	1.10	1.82
22	BMORI:DNA ligase 1-like isoform X2	HaOG206374	1.43	1.81
23	DMELA:Q9VSH9 UPF0183 protein CG7083 BMORI:UPF0183 protein CG7083-like	HaOG213196	2.00	1.77
24	DMELA:Q9VN93 Putative cysteine proteinase CG12163 BMORI:fibroinase precursor	HaOG207683	2.13	1.73
25	DMELA:Q2PDP3 CG2201, isoform E	HaOG204816	2.17	1.70
26	DMELA:Q7KTG2 CG15828, isoform B BMORI:apolipophorins-like	HaOG213763	1.66	1.65
27	DMELA:Q8IMW9 CG6454, isoform C BMORI:C2 domain-containing protein 5-like	HaOG215054	1.92	1.62
28	DMELA:Q8MRN2 GH13342p BMORI:uncharacterized protein LOC101735388	HaOG204554	1.42	1.61
29	DMELA:O76906 Protein cramped BMORI:protein cramped-like	HaOG208924	2.35	1.60
30	DMELA:Q9XZ56 4E-binding protein THOR	HaOG209464	1.98	1.60
31	DMELA:Q9VJE5 Restin homolog BMORI:restin homolog	HaOG212175	1.71	1.54
32	BMORI:xaa-Pro dipeptidase-like isoform X3	HaOG203283	1.07	1.54
33	DMELA:Q7KAK2 Farnesyl pyrophosphate synthase (Dimethylallyltransferase)	HaOG206975	1.42	1.50
34	DMELA:Q7KTS2 CG1347, isoform B	HaOG208115	1.54	1.45
35	BMORI:maternal protein tudor-like	HaOG206847	1.64	1.45
36	DMELA:P49021 Protein timeless BMORI:timeless	HaOG217123	1.77	1.37
37	serine protease gd-like	HaOG209738	1.52	1.36
38	DMELA:Q8SYN0 CG11638	HaOG201973	1.35	1.30
39	BMORI:LOW QUALITY PROTEIN: 5'-3' exoribonuclease 2 homolog	HaOG204131	1.89	1.27
40	BMORI:E3 ubiquitin-protein ligase MARCH5-like	HaOG214126	1.28	1.25
41	BMORI:uncharacterized protein LOC101739953 isoform X1	HaOG213412	1.73	1.23
42	DMELA:Q9VEX0 Extracellular sulfatase SULF-1 homolog BMORI:extracellular sulfatase SULF-1 homolog	HaOG201436	1.33	1.14
43	DMELA:A1Z8F4 Schnurri, isoform D BMORI:uncharacterized protein LOC101737452	HaOG206586	0.94	1.10

44	HSAPI:Q9BQP7 Mitochondrial genome maintenance exonuclease 1 BMORI:mitochondrial genome maintenance exonuclease 1-like isoform X2	HaOG203348	1.57	1.05
45	DMELA:Q9VK36 CG5204	HaOG208091	1.11	1.00
46	DMELA:Q8IGL2 RE69804p BMORI:DnaJ (Hsp40) homolog 9	HaOG207356	0.85	0.97
47	HSAPI:Q16526 Cryptochrome-1 BMORI:cryptochrome 2	HaOG201315	1.49	0.93
48	DMELA:Q9VM33 Elongation factor G, mitochondrial BMORI:elongation factor G, mitochondrial-like	HaOG208027	-1.39	-0.91
49	DMELA:Q7KLW8 LD03471p BMORI:protein SEC13 homolog	HaOG202960	-1.54	-0.94
50	DMELA:P55035 26S proteasome non-ATPase regulatory subunit 4 BMORI:proteasome 26S non-ATPase subunit 4	HaOG211398	-1.19	-0.98
51	DMELA:Q9VPQ2 CG4164	HaOG213230	-1.12	-0.98
52	DMELA:Q9W1V3 rRNA 2'-O-methyltransferase fibrillarin BMORI:rRNA 2'-O-methyltransferase fibrillarin-like	HaOG209384	-1.80	-0.99
53	DMELA:P56538 Eukaryotic translation initiation factor 6 BMORI:eukaryotic translation initiation factor 6	HaOG209996	-1.27	-1.01
54	HSAPI:P28066 Proteasome subunit alpha type-5 BMORI:proteasome zeta subunit	HaOG214956	-1.04	-1.01
55	DMELA:Q9I7K5 Transmembrane emp24 domain-containing protein eca BMORI:transmembrane emp24 protein transport domain containing 9 precursor	HaOG202530	-0.78	-1.02
56	DMELA:Q7K0Y1 Ribosome biogenesis protein BOP1 homolog BMORI:ribosome biogenesis protein BOP1 homolog	HaOG211250	-1.37	-1.03
57	DMELA:Q9VZ23 GTP-binding nuclear protein Ran BMORI:GTP-binding nuclear protein Ran	HaOG212172	-1.71	-1.06
58	DMELA:Q9VX98 CG9099	HaOG212651	-1.78	-1.06
59	BMORI:peptidyl-prolyl cis-trans isomerase	HaOG206578	-1.24	-1.06
60	DMELA:A1Z9E3 Elongation factor Tu BMORI:elongation factor Tu	HaOG212171	-1.83	-1.07
61	DMELA:Q9VXP3 GH05406p	HaOG215666	-1.52	-1.07
62	DMELA:P40796 La protein homolog BMORI:la protein homolog isoform X2	HaOG214975	-2.29	-1.08
63	CELEG:Q9GUM1 Protein Y73E7A.1, isoform a BMORI:coiled-coil domain-containing protein 124-like isoform X3	HaOG209881	-1.93	-1.09
64	DMELA:P35128 Ubiquitin-conjugating enzyme E2 N BMORI:ubiquitin conjugating enzyme E2	HaOG204404	-1.10	-1.10
65	DMELA:P08985 Histone H2A.v BMORI:H2A histone family member V	HaOG211144	-1.23	-1.10
66	HSAPI:P09661 U2 small nuclear ribonucleoprotein A' BMORI:U2 small nuclear ribonucleoprotein A'	HaOG201835	-1.60	-1.10
67	DMELA:Q8SYL1 Lethal (2) 09851	HaOG205531	-2.09	-1.12
68	DMELA:Q9VL78 FK506-binding protein 59 BMORI:FK506-binding protein FKBP59 homologue	HaOG209865	-1.64	-1.13
69	DMELA:Q9VKQ3 Ribosome biogenesis protein WDR12 homolog BMORI:ribosome biogenesis protein WDR12 homolog	HaOG216297	-2.23	-1.14
70	DMELA:P17917 Proliferating cell nuclear antigen BMORI:proliferating cell nuclear antigen	HaOG206623	-1.69	-1.15
71	HSAPI:P54136 Arginine–tRNA ligase, cytoplasmic BMORI:arginine–tRNA ligase, cytoplasmic-like	HaOG202143	-2.14	-1.16
72	DMELA:Q8MR62 Viral IAP-associated factor homolog BMORI:viral IAP-associated factor homolog	HaOG204124	-1.78	-1.16
73	DMELA:A1Z7K8 CG8235 BMORI:endothelial-monocyte activating polypeptide II	HaOG205829	-1.23	-1.19
74	DMELA:Q9VJ38 39S ribosomal protein L13, mitochondrial BMORI:39S ribosomal protein L13, mitochondrial-like	HaOG208604	-1.43	-1.20
75	DMELA:Q7JVG6 CG1550	HaOG213793	-1.65	-1.20
76	DMELA:Q9VXB5 39S ribosomal protein L22, mitochondrial BMORI:39S ribosomal protein L22, mitochondrial-like	HaOG211549	-2.12	-1.21
77	BMORI:uncharacterized protein LOC101740936	HaOG215261	-1.48	-1.21
78	DMELA:Q8IGT5 RE33426p BMORI:nucleolar protein 56-like	HaOG203696	-1.23	-1.22

79	DMELA:Q9VLK2 Ribosomal L1 domain-containing protein CG13096 BMORI:ribosomal L1 domain-containing protein CG13096-like	HaOG209092	-2.21	-1.22
80	DMELA:P41375 Eukaryotic translation initiation factor 2 subunit 2 BMORI:eukaryotic translation initiation factor 2 subunit 2	HaOG215665	-2.05	-1.23
81	DMELA:Q8SZL6 RH10688p BMORI:malectin-like	HaOG205040	-1.60	-1.24
82	DMELA:A1Z6P3 Eb1, isoform F	HaOG202795	-1.06	-1.25
83	HSAPI:Q9BTM9 Ubiquitin-related modifier 1 homolog BMORI:ubiquitin-related modifier 1 homolog	HaOG210774	-1.90	-1.27
84	DMELA:P48592 Ribonucleoside-diphosphate reductase subunit M2 BMORI:ribonucleoside diphosphate reductase small subunit	HaOG211812	-1.45	-1.27
85	DMELA:Q9XZT6 Deoxynucleoside kinase BMORI:putative deoxynucleoside kinase	HaOG204372	-1.75	-1.30
86	DMELA:P43332 U1 small nuclear ribonucleoprotein A BMORI:U1 small nuclear ribonucleoprotein A	HaOG214422	-2.25	-1.40
87	DMELA:Q9VU36 LP04985p	HaOG206747	-1.73	-1.41
88	DMELA:Q9GQ89 Eukaryotic initiation factor eIF2B alpha subunit BMORI:eIF2B-alpha protein	HaOG203919	-1.52	-1.45
89	DMELA:P54622 Single-stranded DNA-binding protein, mitochondrial BMORI:mitochondrial single-stranded DNA-binding protein	HaOG204078	-1.72	-1.47
90	BMORI:elongation of very long chain fatty acids protein 2-like	HaOG206659	-1.79	-1.49
91	DMELA:Q7PLT4 Tim17b BMORI:mitochondrial import inner membrane translocase	HaOG216752	-2.58	-1.59
92	DMELA:Q9VA81 IP02765p	HaOG202483	-1.29	-1.74
93	BMORI:short-chain dehydrogenease/reductase-like	HaOG206177	-1.79	-1.75
94	DMELA:Q9U3Z7 NHP2-like protein 1 homolog BMORI:ribosomal protein L7Ae	HaOG217260	-2.52	-1.76
95	HSAPI:Q9Y3E5 Peptidyl-tRNA hydrolase 2, mitochondrial BMORI:peptidyl-tRNA hydrolase 2, mitochondrial-like	HaOG210038	-2.36	-1.77
96	BMORI:vitellogenin precursor	HaOG208268	-2.31	-1.84
97	BMORI:uncharacterized protein LOC101746705 isoform X1	HaOG209938	-1.91	-1.89
98	DMELA:Q8MKK1 CG30185	HaOG202790	-1.82	-1.94
99	DMELA:Q8T9B2 SD09147p BMORI:mitochondrial ribosomal protein L54	HaOG209676	-1.66	-1.97
100	BMORI:uncharacterized protein LOC101746539	HaOG210604	-3.54	-2.48
	At 12 hours (n=21)			
1	BMORI:p270	HaOG209733	2.70	2.52
2	HaLipase61	HaOG200588	2.10	2.12
3	BMORI:trans-1,2-dihydrobenzene-1,2-diol dehydrogenase-like isoform X2	HaOG211544	2.61	2.02
4	DMELA:Q7KTG2 CG15828, isoform B BMORI:apolipophorins-like	HaOG213763	1.76	1.90
5	BMORI:fumarylacetoacetase	HaOG215692	2.69	1.85
6	CYP4G8-Ha	HaOG200074	1.88	1.76
7	DMELA:P40320 S-adenosylmethionine synthase BMORI:S-adenosylmethionine synthetase	HaOG206272	1.39	1.70
8	BMORI:fatty acid synthase-like	HaOG207602	2.42	1.70
9	DMELA:Q961W1 4-hydroxyphenylpyruvate dioxygenase BMORI:4-hydroxyphenylpyruvate dioxygenase-like	HaOG211200	2.55	1.66
10	DMELA:Q9VA02 Probable 2-oxoglutarate dehydrogenase E1 component DHKTD1 homolog, mitochondrial BMORI:probable 2-oxoglutarate dehydrogenase E1 component DHKTD1 homolog, mitochondrial-like isoform X1	HaOG217047	1.56	1.58
11	HSAPI:Q14624 Inter-alpha-trypsin inhibitor heavy chain H4 BMORI:inter-alpha-trypsin inhibitor heavy chain H4-like	HaOG201450	2.00	1.54
12	DMELA:Q7KAK2 Farnesyl pyrophosphate synthase (Dimethylallyltransferase)	HaOG206975	1.23	1.53
13	DMELA:Q9VSY0 Cuticular protein 67B	HaOG206673	1.64	1.47
14	serine protease gd-like	HaOG209738	1.93	1.43

15	DMELA:A1Z992 CG33138	HaOG204678	1.46	1.40
16	BMORI:poly(U)-specific endoribonuclease homolog	HaOG206610	1.50	1.35
17	DMELA:P17336 Catalase BMORI:catalase	HaOG215508	1.08	1.11
18	HSAPI:P54136 Arginine–tRNA ligase, cytoplasmic BMORI:arginine–tRNA ligase, cytoplasmic-like	HaOG202143	-1.07	-0.96
19	DMELA:Q9VJ38 39S ribosomal protein L13, mitochondrial BMORI:39S ribosomal protein L13, mitochondrial-like	HaOG208604	-1.32	-1.16
20	DMELA:Q7JVK1 HL07956p	HaOG208694	-1.14	-1.32
21	DMELA:Q9VXB5 39S ribosomal protein L22, mitochondrial BMORI:39S ribosomal protein L22, mitochondrial-like	HaOG211549	-1.28	-1.45
	At 24 hours (n=89)			
1	BMORI:cuticular protein RR-1 motif 46 precursor	HaOG201905	3.49	3.98
2	BMORI:cuticular protein RR-1 motif 46 precursor	HaOG201907	4.57	3.79
3	BMORI:cuticular protein RR-1 motif 46 precursor	HaOG201908	3.93	3.72
4	HarmCSP23.1	HaOG200660	2.00	3.63
5	BMORI:p260	HaOG204585	2.96	3.54
6	BMORI:fatty acid synthase-like	HaOG207602	2.83	3.43
7	HarUGT33J1 ALT:HarUGT-23	HaOG200284	4.22	3.34
8	BMORI:juvenile hormone binding protein an-0921 precursor	HaOG204351	4.40	3.22
9	BMORI:acyl-CoA Delta(11) desaturase	HaOG216651	2.73	3.17
10	BMORI:vitellogenin precursor	HaOG208268	2.83	3.16
11	DMELA:Q961W1 4-hydroxyphenylpyruvate dioxygenase BMORI:4-hydroxyphenylpyruvate dioxygenase-like	HaOG211200	4.16	3.09
12	CYP4AU7-Ha-26	HaOG200069	2.97	3.06
13	CYP4G9-Ha	HaOG200075	1.92	3.00
14	BMORI:solute carrier family 46 member 3-like	HaOG206634	1.69	2.95
15	BMORI:alpha-tocopherol transfer protein-like	HaOG210356	2.59	2.89
16	0	HaOG204395	3.57	2.85
17	HarUGT40M1 ALT:HarUGT-14	HaOG200264	4.22	2.75
18	BMORI:nuclear factor of activated T-cells 5-like	HaOG206901	3.79	2.60
19	DMELA:Q9W5X1 CG9572	HaOG207458	2.78	2.57
20	HaLipase61	HaOG200588	2.75	2.53
21	BMORI:phosphoribosyl pyrophosphate synthetase	HaOG202408	1.76	2.37
22	BMORI:trans-1,2-dihydrobenzene-1,2-diol dehydrogenase-like isoform X2	HaOG211544	2.52	2.32
23	DMELA:Q9XTP7 DOMINA protein	HaOG216334	2.30	2.24
24	BMORI:retinal dehydrogenase 1-like	HaOG216699	1.72	2.21
25	CYP6AE19-Ha	HaOG200094	1.86	2.14
26	HarmOBP6	HaOG200803	2.21	2.13
27	BMORI:cuticular protein glycine-rich 10 precursor	HaOG215934	1.61	2.08
28	BMORI:uncharacterized protein LOC101737025	HaOG213700	1.53	2.08
29	HaABCG1 ALT:HaABC-G-05-1-H	HaOG200346	2.42	2.05
30	CYP4G26-Ha	HaOG200072	2.08	2.03
31	HSAPI:Q3LXA3 Bifunctional ATP-dependent dihydroxyacetone kinase/FAD-AMP lyase (cyclizing) BMORI:bifunctional ATP-dependent dihydroxyacetone kinase/FAD-AMP lyase (cyclizing)-like	HaOG213660	1.47	1.96
32	BMORI:C-type lectin 10 precursor	HaOG205089	3.56	1.94
33	BMORI:LOW QUALITY PROTEIN: aldo-keto reductase family 1 member C4-like	HaOG203859	2.40	1.91
34	BMORI:cuticular protein RR-1 motif 42 precursor	HaOG201900	1.59	1.87
35	BMORI:cuticular protein RR-1 motif 11 precursor	HaOG202657	1.59	1.83
36	DMELA:Q9VSY0 Cuticular protein 67B	HaOG206673	1.80	1.82
37	HaGSTe16	HaOG200231	2.37	1.82
38	HSAPI:Q6P2I3 Fumarylacetoacetate hydrolase domain-containing protein 2B BMORI:fumarylacetoacetase	HaOG215554	2.13	1.76
39	BMORI:von Willebrand factor A domain-containing protein 8-like	HaOG204383	2.14	1.74
40	BMORI:putative hexokinase HKDC1-like	HaOG212535	1.71	1.74
41	DMELA:Q9VQB4 CG3609	HaOG204750	2.06	1.73

42	DMELA:P46415 Alcohol dehydrogenase class-3 BMORI:alcohol dehydrogenase	HaOG217257	2.47	1.69
43	BMORI:nitric oxide synthase	HaOG206412	1.55	1.64
44	BMORI:uncharacterized protein LOC101746794	HaOG213603	2.55	1.63
45	BMORI:antichymotrypsin-1 precursor	HaOG207953	1.99	1.63
46	DMELA:Q962N6 Flavin-containing monooxygenase FMO-1 BMORI:flavin-dependent monooxygenase FMO2 precursor	HaOG207733	1.68	1.61
47	BMORI:juvenile hormone binding protein an-0128 precursor	HaOG204844	1.68	1.60
48	BMORI:uncharacterized protein LOC101744418	HaOG202396	1.53	1.60
49	BMORI:leucine-rich repeat-containing protein 15-like	HaOG211658	1.51	1.56
50	BMORI:putative fatty acyl-CoA reductase CG5065-like	HaOG210892	1.32	1.47
51	DMELA:Q9V4B8 CG31999	HaOG211690	1.58	1.46
52	DMELA:Q9XZ56 4E-binding protein THOR	HaOG209464	1.49	1.44
53	CYP333A1-Ha	HaOG200023	1.86	1.42
54	BMORI:elongation of very long chain fatty acids protein AAEL008004-like	HaOG206661	2.60	1.37
55	BMORI:elongation of very long chain fatty acids protein AAEL008004-like	HaOG206662	1.83	1.33
56	DMELA:Q7K2E1 CG8839, isoform A	HaOG207822	1.81	1.30
57	serine proteinase-like protein precursor	HaOG211226	1.49	1.20
58	DMELA:A1ZBF3 Juvenile hormone epoxide hydrolase 2, isoform C	HaOG214591	1.12	1.15
59	DMELA:A1Z6Z3 Aldehyde dehydrogenase BMORI:aldehyde dehydrogenase isoform 1	HaOG205660	0.96	1.11
60	BMORI:zinc finger protein 600-like	HaOG215539	1.48	1.06
61	DMELA:Q7KAK2 Farnesyl pyrophosphate synthase (Dimethylallyltransferase)	HaOG206975	1.35	1.06
62	DMELA:Q8MQS7 Malate dehydrogenase BMORI:cytosolic malate dehydrogenase	HaOG205283	0.88	1.02
63	DMELA:Q3YNC0 CG9847 BMORI:FK506-binding protein precursor	HaOG215867	-1.28	-1.09
64	DMELA:Q7JRJ1 Ell-associated factor Eaf BMORI:ELL-associated factor 2-like	HaOG202866	-1.17	-1.18
65	DMELA:E1JHE4 Fatty acid (Long chain) transport protein, isoform C BMORI:fatty acid transport protein	HaOG210925	-0.98	-1.23
66	DMELA:A1Z8N1 Facilitated trehalose transporter Tret1-1 BMORI:facilitated trehalose transporter Tret1	HaOG215283	-1.18	-1.32
67	DMELA:Q8IPN4 Krotzkopf verkehrt, isoform A	HaOG210462	-1.38	-1.38
68	DMELA:E1JI40 Vermiform, isoform E	HaOG217147	-1.79	-1.57
69	DMELA:Q9VQH2 Dual oxidase BMORI:dual oxidase-like	HaOG212627	-2.55	-1.62
70	DMELA:Q9VZV2 Cht7	HaOG211353	-1.62	-1.70
71	DMELA:Q8SY08 RE18374p BMORI:decaprenyl-diphosphate synthase subunit 1-like, partial	HaOG206224	-1.92	-1.78
72	DMELA:Q9VAN1 CG14515 BMORI:neuropeptide receptor B3 precursor	HaOG206273	-1.64	-1.78
73	BMORI:gamma-interferon-inducible lysosomal thiol reductase-like	HaOG215278	-1.85	-1.93
74	BMORI:mucin-17-like	HaOG215196	-2.29	-2.00
75	BMORI:uncharacterized protein LOC101739142	HaOG202357	-2.55	-2.06
76	BMORI:mucin-5AC-like isoform X5	HaOG214737	-2.73	-2.07
77	DMELA:C0PTW4 CG8776, isoform F	HaOG205076	-1.83	-2.12
78	BMORI:uncharacterized protein LOC101737216	HaOG206589	-1.71	-2.12
79	DMELA:Q9VR79 LD43683p	HaOG206891	-2.83	-2.14
80	BMORI:neurogenic locus notch homolog protein 3-like	HaOG212397	-2.46	-2.14
81	BMORI:myristoylated alanine-rich C-kinase substrate-like isoform X1	HaOG214739	-2.07	-2.31
82	DMELA:A8JQU3 Gasp, isoform B BMORI:gasp precursor	HaOG206885	-2.68	-2.34
83	DMELA:Q8T0J5 CG7675, isoform A	HaOG203692	-2.08	-2.42
84	BMORI:uncharacterized protein LOC101739160	HaOG205610	-2.31	-2.48
85	DMELA:Q0KIA5 Mind the gap BMORI:uncharacterized protein LOC101736353	HaOG215047	-2.42	-2.48

86	BMORI:collagen alpha-1(II) chain-like	HaOG215487	-2.69	-3.03
87	DMELA:Q8I0P8 Cuticular protein 65Av	HaOG214287	-1.73	-3.14
88	BMORI:alpha-tocopherol transfer protein-like	HaOG210360	-2.25	-3.77
89	BMORI:cuticular protein RR-1 motif 32 precursor	HaOG201888	-3.41	-4.10

Table B.1: List of DE genes that are present in both UU and SS strains when comparing between exposed an unexposed cohorts within a strain. There are 100, 21 and 89 DE genes at the 6, 12 and 24-hour timepoints respectively. Genes are ranked by highest level of differential expression in the SS strain. The level of differential expression for each strain is shown as $\log_2$ fold-change values. Negative values represent downregulation in the exposed cohort relative to the unexposed cohort for a strain.

Appendix C

List of 219 genes that are differentially expressed between treatments and between strains

	HaOG	Gene annotation
1	HaOG216548	BMORI:27 kDa glycoprotein precursor
2	HaOG209773	BMORI:acyl-CoA desaturase
3	HaOG211959	BMORI:aldose reductase-like
4	HaOG202369	BMORI:aliphatic nitrilase
5	HaOG210360	BMORI:alpha-tocopherol transfer protein-like
6	HaOG212906	BMORI:ankyrin repeat and sterile alpha motif domain-containing protein 1B-like
7	HaOG207953	BMORI:antichymotrypsin-1 precursor
8	HaOG207595	BMORI:apolipoprotein D-like
9	HaOG213884	BMORI:AT-rich interactive domain-containing protein 4B-like
10	HaOG205154	BMORI:calponin homology domain-containing protein DDB-G0272472-like
11	HaOG212060	BMORI:cellular retinaldehyde-binding protein
12	HaOG215601	BMORI:cuticular protein glycine-rich 13 precursor
13	HaOG210770	BMORI:cuticular protein glycine-rich 6 precursor
14	HaOG201905	BMORI:cuticular protein RR-1 motif 46 precursor
15	HaOG212663	BMORI:cysteine-rich with EGF-like domain protein 1-like
16	HaOG212897	BMORI:cytoplasmic dynein 2 heavy chain 1-like
17	HaOG206659	BMORI:elongation of very long chain fatty acids protein 2-like
18	HaOG206662	BMORI:elongation of very long chain fatty acids protein AAEL008004-like
19	HaOG210112	BMORI:fatty acid-binding protein, heart-like
20	HaOG215346	BMORI:fibroin light chain precursor
21	HaOG201542	BMORI:flocculation protein FLO11-like
22	HaOG204363	BMORI:hemolymph juvenile hormone binding protein precursor
23	HaOG204364	BMORI:hemolymph juvenile hormone binding protein precursor
24	HaOG217234	BMORI:laminin subunit alpha-like
25	HaOG206694	BMORI:leucine-rich repeat extensin-like protein 3-like
26	HaOG208708	BMORI:luciferin 4-monooxygenase-like
27	HaOG206815	BMORI:monocarboxylate transporter 14-like, partial
28	HaOG215196	BMORI:mucin-17-like
29	HaOG212338	BMORI:mucin-2-like
30	HaOG214737	BMORI:mucin-5AC-like isoform X5
31	HaOG203373	BMORI:myosin-11-like
32	HaOG214739	BMORI:myristoylated alanine-rich C-kinase substrate-like isoform X1
33	HaOG212206	BMORI:neurogenic locus notch homolog protein 1-like
34	HaOG210592	BMORI:neuroglian-like
35	HaOG206412	BMORI:nitric oxide synthase
36	HaOG204870	BMORI:organic cation transporter protein-like
37	HaOG202408	BMORI:phosphoribosyl pyrophosphate synthetase
38	HaOG212161	BMORI:probable ATP-dependent RNA helicase ddx17-like
39	HaOG203426	BMORI:probable pseudouridine-5'-monophosphatase-like isoform X1
40	HaOG207975	BMORI:protein ECT2-like
41	HaOG206635	BMORI:proton-coupled folate transporter-like
42	HaOG211045	BMORI:putative transcription factor SOX-15-like
43	HaOG211451	BMORI:quinone oxidoreductase-like
44	HaOG216699	BMORI:retinal dehydrogenase 1-like
45	HaOG202196	BMORI:RNA exonuclease 1 homolog
46	HaOG206634	BMORI:solute carrier family 46 member 3-like
47	HaOG205586	BMORI:synaptic vesicle glycoprotein 2B-like isoform X1
48	HaOG213725	BMORI:trichohyalin-like
49	HaOG202528	BMORI:U3 small nucleolar RNA-associated protein 14 homolog A-like
50	HaOG201276	BMORI:uncharacterized protein LOC101736300, partial
51	HaOG206589	BMORI:uncharacterized protein LOC101737216
52	HaOG209299	BMORI:uncharacterized protein LOC101738032
53	HaOG204753	BMORI:uncharacterized protein LOC101740603
54	HaOG210344	BMORI:uncharacterized protein LOC101740930
55	HaOG212621	BMORI:uncharacterized protein LOC101741832 isoform X2
56	HaOG215502	BMORI:uncharacterized protein LOC101742068
57	HaOG207617	BMORI:uncharacterized protein LOC101742291
58	HaOG212317	BMORI:uncharacterized protein LOC101742721

59	HaOG209335	BMORI:uncharacterized protein LOC101744544
60	HaOG206713	BMORI:uncharacterized protein LOC101745021
61	HaOG206325	BMORI:uncharacterized protein LOC101745862
62	HaOG214825	BMORI:uncharacterized protein LOC101746940
63	HaOG202169	BMORI:vegetative cell wall protein gp1-like
64	HaOG205395	BMORI:VID27-like protein-like
65	HaOG208266	BMORI:vitellogenin precursor
66	HaOG209212	BMORI:yellow-b precursor
67	HaOG202098	BMORI:zinc finger protein 184-like
68	HaOG202619	BMORI:zinc finger protein 791-like
69	HaOG211452	CELEG:O45496 Protein F39B2.3 BMORI:quinone oxidoreductase-like
70	HaOG205635	CELEG:Y69H2.3c.1 NULL NULL BMORI:tenascin-like
71	HaOG200023	CYP333A1-Ha
72	HaOG200072	CYP4G26-Ha
73	HaOG200094	CYP6AE19-Ha
74	HaOG200098	CYP6AN1-Ha
75	HaOG200099	CYP6AN3-Ha
76	HaOG216172	DMELA:A1Z6K0 CG7849, isoform A
77	HaOG209351	DMELA:A1Z6X6 CG1707
78	HaOG209622	DMELA:A1Z746 Kynurenine 3-monooxygenase BMORI:kynurenine 3-monooxygenase
79	HaOG201384	DMELA:A1Z8D2 CG18003, isoform A
80	HaOG215283	DMELA:A1Z8N1 Facilitated trehalose transporter Tret1-1 BMORI:facilitated trehalose transporter Tret1
81	HaOG212787	DMELA:A1Z8R0 CG13188, isoform B
82	HaOG205965	DMELA:A1Z942 No extended memory, isoform E
83	HaOG209146	DMELA:A1ZA45 CG30089 BMORI:uncharacterized protein LOC101744784
84	HaOG215908	DMELA:A8DYS5 Chitin deacetylase-like 5, isoform D
85	HaOG217198	DMELA:A8E6R2 CG11241, isoform B
86	HaOG206885	DMELA:A8JQU3 Gasp, isoform B BMORI:gasp precursor
87	HaOG211871	DMELA:B5RJS0 CG42388, isoform E
88	HaOG203738	DMELA:B7YZT7 CG8740, isoform D BMORI:dentin sialophosphoprotein-like
89	HaOG215412	DMELA:B7Z097 CG32442, isoform C BMORI:uncharacterized protein LOC101740229
90	HaOG205006	DMELA:B7Z0Q3 CG11857, isoform B BMORI:protein RER1-like isoform X1
91	HaOG217147	DMELA:E1JI40 Vermiform, isoform E
92	HaOG210448	DMELA:O15943 Neural-cadherin BMORI:neural-cadherin
93	HaOG208956	DMELA:O18407 Collagen type IV alpha 2 BMORI:LOW QUALITY PROTEIN: collagen alpha-2(IV) chain-like
94	HaOG216708	DMELA:O18660 BZIP transcription factor BMORI:nuclear factor interleukin-3-regulated protein-like
95	HaOG215902	DMELA:O77459 Transcription factor Ken BMORI:ken and barbie protein
96	HaOG208116	DMELA:P12646 Glucose-6-phosphate 1-dehydrogenase BMORI:glucose-6-phosphate 1-dehydrogenase-like
97	HaOG206623	DMELA:P17917 Proliferating cell nuclear antigen BMORI:proliferating cell nuclear antigen
98	HaOG204787	DMELA:P18459 Tyrosine 3-monooxygenase BMORI:tyrosine hydroxylase
99	HaOG214709	DMELA:P33244 Nuclear hormone receptor FTZ-F1 BMORI:nuclear hormone receptor FTZ-F1
100	HaOG209799	DMELA:P48591 Ribonucleoside-diphosphate reductase large subunit BMORI:ribonucleoside-diphosphate reductase large subunit-like
101	HaOG205935	DMELA:P48608 Protein diaphanous BMORI:protein diaphanous-like
102	HaOG201798	DMELA:P81928 RPII140-upstream gene protein BMORI:RPII140-upstream gene protein-like
103	HaOG208619	DMELA:P82147 Protein lethal(2)essential for life BMORI:heat shock protein 1
104	HaOG202552	DMELA:Q03043 cGMP-dependent protein kinase, isozyme 2 forms cD4/T1/T3A/T3B BMORI:protein kinase, cGMP-dependent, type I
105	HaOG205083	DMELA:Q0E8W1 CG30427, isoform A
106	HaOG208526	DMELA:Q0E908 Hillarin, isoform A
107	HaOG215255	DMELA:Q0E960 CG8389, isoform A
108	HaOG215047	DMELA:Q0KIA5 Mind the gap BMORI:uncharacterized protein LOC101736353
109	HaOG216188	DMELA:Q24492 Replication protein A 70 kDa DNA-binding subunit BMORI:replication protein A1

110	HaOG206413	DMELA:Q27571 Nitric oxide synthase BMORI:nitric oxide synthase-like protein
111	HaOG204816	DMELA:Q2PDP3 CG2201, isoform E
112	HaOG212842	DMELA:Q5BID6 RE32747p BMORI:short-chain dehydrogenase/reductase family 16C member 6-like
113	HaOG209280	DMELA:Q5BIH5 CG12753, isoform B
114	HaOG211384	DMELA:Q7JR58 CG6543, isoform A
115	HaOG201252	DMELA:Q7JWM6 CG10936, isoform A
116	HaOG203137	DMELA:Q7JWQ7 CG3074, isoform A
117	HaOG205825	DMELA:Q7K172 LD04933p
118	HaOG207427	DMELA:Q7K2V9 CG33671
119	HaOG205510	DMELA:Q7K537 GH14316p BMORI:selenium-binding protein 1-like isoform X1
120	HaOG205418	DMELA:Q7KV91 Purine nucleoside phosphorylase BMORI:purine nucleoside phosphorylase-like isoform X1
121	HaOG210519	DMELA:Q7KVI9 CG9812, isoform B BMORI:uncharacterized protein LOC101735353
122	HaOG211346	DMELA:Q7YU25 RE05911p BMORI:uncharacterized protein LOC101744777
123	HaOG213834	DMELA:Q86P00 LP10544p BMORI:CCA tRNA nucleotidyltransferase 1, mitochondrial-like
124	HaOG202350	DMELA:Q8IGE6 Ribose-phosphate pyrophosphokinase BMORI:phosphoribosyl pyrophosphate synthetase
125	HaOG209339	DMELA:Q8IH53 GH12864p BMORI:porphobilinogen deaminase-like
126	HaOG209907	DMELA:Q8IMI9 CG31028, isoform D BMORI:uncharacterized protein LOC101739289
127	HaOG207575	DMELA:Q8IND1 Arpc3A, isoform D
128	HaOG210462	DMELA:Q8IPN4 Krotzkopf verkehrt, isoform A
129	HaOG207679	DMELA:Q8IPQ7 AT06280p
130	HaOG216329	DMELA:Q8MQI6 Transcription elongation factor 1 homolog BMORI:transcription elongation factor 1 homolog
131	HaOG212652	DMELA:Q8MQJ7 Autophagy-specific gene 1, isoform B
132	HaOG202726	DMELA:Q8MQN1 RE64786p BMORI:homocysteine S-methyltransferase
133	HaOG209114	DMELA:Q8MRY2 SD14156p BMORI:ubiquitin carboxyl-terminal hydrolase 43-like
134	HaOG205167	DMELA:Q8MSU3 Putative ferric-chelate reductase 1 homolog BMORI:putative ferric-chelate reductase 1 homolog
135	HaOG212948	DMELA:Q8SXH6 RE44714p BMORI:uncharacterized protein LOC101739898
136	HaOG206931	DMELA:Q8SXK3 RE16411p BMORI:2-amino-3-ketobutyrate coenzyme A ligase, mitochondrial-like
137	HaOG206224	DMELA:Q8SY08 RE18374p BMORI:decaprenyl-diphosphate synthase subunit 1-like, partial
138	HaOG209657	DMELA:Q8SY12 RE15159p BMORI:hydroxylysine kinase-like
139	HaOG209889	DMELA:Q8SYC6 RE68083p
140	HaOG201462	DMELA:Q8SYN7 RE50056p BMORI:ribonuclease H2 subunit B-like
141	HaOG205040	DMELA:Q8SZL6 RH10688p BMORI:malectin-like
142	HaOG210667	DMELA:Q8WQM0 Rho GTPase guanine nucleotide exchange factor GEF64C BMORI:uncharacterized protein LOC101738416
143	HaOG205481	DMELA:Q94517 Histone deacetylase Rpd3 BMORI:LOW QUALITY PROTEIN: histone deacetylase Rpd3-like
144	HaOG206619	DMELA:Q94548 LK6 protein kinase BMORI:MAP kinase-interacting serine/threonine-protein kinase 2-like
145	HaOG201499	DMELA:Q95S24 GM14561p BMORI:protein canopy homolog 3-like
146	HaOG207733	DMELA:Q962N6 Flavin-containing monooxygenase FMO-1 BMORI:flavin-dependent monooxygenase FMO2 precursor
147	HaOG214854	DMELA:Q9V447 Protein Kr-h2 BMORI:protein Kr-h2-like
148	HaOG211110	DMELA:Q9V496 Apolipophorins BMORI:apolipophorins isoform X2
149	HaOG211690	DMELA:Q9V4B8 CG31999
150	HaOG217125	DMELA:Q9V6X7 GDP-fucose protein O-fucosyltransferase 1 BMORI:protein-O-fucosyltransferase 1 precursor
151	HaOG207201	DMELA:Q9VB96 CG31075
152	HaOG205860	DMELA:Q9VC61 Protein CREBRF homolog BMORI:uncharacterized protein LOC101740805
153	HaOG202767	DMELA:Q9VCD0 Glutamyl-tRNA(Gln) amidotransferase subunit B, mitochondrial BMORI:glutamyl-tRNA amidotransferase subunit B
154	HaOG210891	DMELA:Q9VCF6 CG12268, isoform A
155	HaOG213116	DMELA:Q9VCG4 Nucleoporin Ndc1 BMORI:nucleoporin NDC1-like

156	HaOG216328	DMELA:Q9VCI4 CG10217, isoform A
157	HaOG212905	DMELA:Q9VCM6 CG4393
158	HaOG213494	DMELA:Q9VDI1 LD46328p
159	HaOG207522	DMELA:Q9VIC9 CG8665
160	HaOG216975	DMELA:Q9VJQ3 RH54244p
161	HaOG202389	DMELA:Q9VLC5 Aldehyde dehydrogenase
162	HaOG216270	DMELA:Q9VMC6 CG9547
163	HaOG201232	DMELA:Q9VN27 Putative lipoyltransferase 2, mitochondrial BMORI:putative lipoyltransferase 2, mitochondrial-like
164	HaOG216767	DMELA:Q9VN86 AT14148p
165	HaOG203254	DMELA:Q9VPG1 CG5847 BMORI:proteoglycan 4-like
166	HaOG208065	DMELA:Q9VPH2 DNA primase large subunit BMORI:DNA primase large subunit-like
167	HaOG209463	DMELA:Q9VQ86 CG15385
168	HaOG212627	DMELA:Q9VQH2 Dual oxidase BMORI:dual oxidase-like
169	HaOG206676	DMELA:Q9VQS4 Spindly BMORI:protein Spindly-like
170	HaOG206891	DMELA:Q9VR79 LD43683p
171	HaOG213196	DMELA:Q9VSH9 UPF0183 protein CG7083 BMORI:UPF0183 protein CG7083-like
172	HaOG208803	DMELA:Q9VTZ5 LD22449p
173	HaOG216446	DMELA:Q9VU70 Tetratricopeptide repeat protein 36 homolog BMORI:tetratricopeptide repeat protein 36-like
174	HaOG207574	DMELA:Q9VWF8 CG14210
175	HaOG209182	DMELA:Q9VY87 CG10992
176	HaOG201682	DMELA:Q9VYT0 CG15739
177	HaOG213331	DMELA:Q9VYY2 Signal peptidase complex subunit 2 BMORI:signal peptidase complex subunit 2
178	HaOG211353	DMELA:Q9VZV2 Cht7
179	HaOG205458	DMELA:Q9W0V1 3-phosphoinositide-dependent protein kinase 1 BMORI:3-phosphoinositide-dependent protein kinase 1-like isoform X1
180	HaOG211554	DMELA:Q9W1G0 Probable transaldolase BMORI:transaldolase
181	HaOG217200	DMELA:Q9W1Y1 ER membrane protein complex subunit 8/9 homolog BMORI:ER membrane protein complex subunit 8/9 homolog
182	HaOG205801	DMELA:Q9W3N9 CG10932
183	HaOG210589	DMELA:Q9W3W5 Protein shifted BMORI:Wnt inhibitory factor 1 precursor
184	HaOG210031	DMELA:Q9W3W7 CG14439
185	HaOG216259	DMELA:Q9XZS5 CG17636
186	HaOG200346	HaABCG1 ALT:HaABC-G-05-1-H
187	HaOG200341	HaABCH2 ALT:HaABC-H-26-2-H
188	HaOG200131	HaCCE017
189	HaOG200140	HaCCE026
190	HaOG200189	HaCCE107
191	HaOG200215	HaGSTe02
192	HaOG200231	HaGSTe16
193	HaOG200235	HaGSTo02
194	HaOG200239	HaGSTs01
195	HaOG200248	HaGSTs09
196	HaOG200249	HaGSTs10
197	HaOG200660	HarmCSP23.1
198	HaOG200803	HarmOBP6
199	HaOG200267	HarUGT33T1 ALT:HarUGT-20
200	HaOG200264	HarUGT40M1 ALT:HarUGT-14
201	HaOG202208	HSAPI:B7Z4K4 Putative tRNA (cytidine(32)/guanosine(34)-2'-O)-methyltransferase
202	HaOG209124	HSAPI:O75832 26S proteasome non-ATPase regulatory subunit 10 BMORI:26S proteasome non-ATPase regulatory subunit 10-like
203	HaOG201835	HSAPI:P09661 U2 small nuclear ribonucleoprotein A' BMORI:U2 small nuclear ribonucleoprotein A'
204	HaOG207212	HSAPI:P23786 Carnitine O-palmitoyltransferase 2, mitochondrial BMORI:carnitine O-palmitoyltransferase 2, mitochondrial-like
205	HaOG210790	HSAPI:P31937 3-hydroxyisobutyrate dehydrogenase, mitochondrial BMORI:3-hydroxyisobutyrate dehydrogenase
206	HaOG212048	HSAPI:P36959 GMP reductase 1 BMORI:GMP reductase 1-like
207	HaOG209379	HSAPI:P54105 Methylosome subunit pICln BMORI:methylosome subunit pICln-like

208	HaOG202122	HSAPI:Q6MZP7 Protein lin-54 homolog BMORI:protein lin-54 homolog
209	HaOG215618	HSAPI:Q6ZNB7 Alkylglycerol monooxygenase BMORI:alkylglycerol monooxygenase-like
210	HaOG216695	HSAPI:Q8NBP5 Major facilitator superfamily domain-containing protein 9 BMORI:major facilitator superfamily domain-containing protein 9-like isoform X1
211	HaOG212955	HSAPI:Q8NHV4 Protein NEDD1 BMORI:uncharacterized protein LOC101745951
212	HaOG206211	HSAPI:Q8WXX5 DnaJ homolog subfamily C member 9 BMORI:DnaJ (Hsp40) homolog 10
213	HaOG203304	HSAPI:Q92673 Sortilin-related receptor BMORI:sortilin-related receptor-like
214	HaOG206851	HSAPI:Q96G03 Phosphoglucomutase-2 BMORI:phosphoglucomutase-2-like
215	HaOG208272	HSAPI:Q9H334 Forkhead box protein P1 BMORI:forkhead box protein P1-like
216	HaOG203554	HSAPI:Q9UHK6 Alpha-methylacyl-CoA racemase BMORI:alpha-methylacyl-CoA racemase-like isoform X1
217	HaOG211579	HSAPI:Q9UNS2 COP9 signalosome complex subunit 3 BMORI:COP9 signalosome complex subunit 3-like
218	HaOG202754	no annotation
219	HaOG214947	no annotation

Table C.1: List of genes that are differentially expressed between treatments and between strains (n=219)